Behavioral Drivers of Routing Decisions: Evidence from Restaurant Table Assignment

Tom F. Tan

Southern Methodist University

6212 Bishop Boulevard Dallas, TX 75275

Tel: 214.768.1228

ttan@cox.smu.edu

Bradley R. Staats

University of North Carolina at Chapel Hill

Campus Box 3490, McColl Building Chapel Hill, NC 27599-3490

Tel: 919.962.7343

bstaats@unc.edu

September 30, 2019

Abstract

We first theoretically identify the factors that may impact individuals' routing decisions before empirically examining a large operational dataset in a casual restaurant setting. Analytical models have identified various routing algorithms for service operations management. Although each model may offer advantages over others, they all make a key assumption - decision makers will actually follow the algorithms, if implemented. However, in many settings routing is not done by a computer that is programmed, but instead by a human. People make routing decisions at their own discretion which may hurt or help system performance. We analyze granular transaction data to examine how hosts revise a given routing rule when seating customers. After that, we empirically analyze the effect of the dispersion of table assignments on restaurant performance, and estimate the counter-factual sales impact of adopting an alternative routing priority. Our setting instructs its hosts to follow a round-robin rule to assign tasks because it ensures fairness and smooths work flow. We find that hosts assign more incoming parties than the round-robin rule suggests to those waiters who have low contemporaneous workload or high speed skills. The prioritization of high speed skill waiters increases with higher levels of demand. In addition, we show an inverted-U-shaped relationship between the inequality of table assignments (measured in terms of the Gini Coefficient of the numbers of tables assigned to each waiter during the same hour) and total sales. Our results suggest that properly adjusting the round-robin rule is productive; however, too much deviation lowers performance. Our paper empirically highlights the value of routing decisions and front-line personnel, such as the hosts in our context.

Keywords: behavioral routing decisions; quality/speed trade-off; fairness; workload; business analytics; restaurant operations; service operations; behavioral operations management

1 Introduction

When designing a service system with multiple servers, a manager has a key question to answer - how should she assign the work to different servers? For example, call centers need to decide which agent should answer an incoming call (Gans et al., 2003; Akşin et al., 2007), hospitals must allocate incoming patients from emergency department into internal wards (Mandelbaum et al., 2012), emergency medical services agencies need to direct ambulances to hospitals (Deo and Gurvich, 2011), and airlines have to decide which maintenance station to route their aircraft for service (Gopalan and Talluri, 1998). The routing problem is a difficult one because agents are heterogeneous in their skill sets and capacities, and because service time and request arrivals are stochastic. Despite its difficulty, solving the routing problem has important implications for companies to enhance service efficiency and quality, generate worker and customer satisfaction, and gain competitive advantage.

Because of its practical significance, the development of routing algorithms or heuristics has generated wide interest in academic circles (e.g., Gans et al., 2003; see also Akşin et al., 2007). However, most research on routing has been devoted to analytical models, as opposed to empirical examination. Insightful as the analytical models are, they assume that managers will strictly follow these algorithms/heuristics if they are implemented as routing rules. However, humans are not machines. They are susceptible to their own incentives and heuristics, which may make them deviate from the given rule (e.g., Schweitzer and Cachon, 2000; Van Donselaar et al., 2010; Cui et al., 2015; Secchi et al., 2019). Such deviation may hurt operational performance (Tucker, 2015), or alternatively, it may improve operational outcomes if frontline workers possess relevant information that is not accounted for in the implemented rule (Bowman, 1963). Therefore, it is important to first learn whether or not individuals actually adjust the given routing rule in an applied setting. It is equally necessary to understand what agent-related factors may influence individuals' routing decisions. In so doing it is then possible to understand how the potential adjustment behavior affects service system performance.

We examine a detailed operational data set about hosts' seating incoming customers at a large full-service American casual restaurant chain. In this setting, hosts are explicitly instructed to follow a round-robin (RR) rule, a commonly used rule[1], to seat customers across waiters in sequence. That is, the host assigns incoming parties by rotation to each waiter, thus equally distributing work among waiters (Walker, 2007, p. 229). We postulate that the host may adjust routing based on three waiter-related factors, which are both theorized in the operations management literature and are empirically testable with our data. They include waiters' workload upon job assignment, their speed skills and sales skills.

To investigate these factors, we analyze two million check-level observations from point-of-sales (POS) data and labor working time data that take place over nine-months (December 2012 to February 2013, December 2013 to February 2014, May 2014 to July 2014) from all 66 restaurants in the chain that are located in one metropolitan area in the southwestern United States. We also shadowed and interviewed hosts, managers and waiters to gain additional institutional knowledge. Controlling for numerous factors, we find that hosts do not strictly follow the RR rule. They assign more tables to waiters that have lower workload or higher speed skills than the RR rule suggests, but they do not prioritize those waiters with higher sales skills, on average. The inequality of the table assignment is moderated by store demand and

[1] The RR rule is the norm in casual full service restaurants to assign customers to waiters. Besides in the restaurants, the RR rule is also used in other service settings, such as emergency departments where triage nurses assign patients to physicians and nurses (Song et al., 2015; Valentine, 2018). Although the RR rule is an important routing method used in the service sector, we are not arguing it is either optimal or generalizable in all service settings.

waiter's average party size. As store demand increases, hosts tend to give more tables to faster waiters. In addition, on average, hosts assign larger parties to waiters having low contemporaneous workload, low speed ability and high sales ability. Furthermore, we find that as the inequality of the table assignments increases (measured in terms of the Gini Coefficient of the number of tables assigned to the waiters during the same hour), total store sales first increase and then decrease. In other words, there is an inverted-U-shaped relationship between the deviation from the RR rule and the total sales. Adjusting the RR rule is productive; however, too much deviation may lower performance. In particular, the optimal Gini Coefficient is approximately 0.4 standard deviation above the sample mean, which suggests that the hosts' adjustments are productive in the right direction, and that additional deviations could aid performance even more. Our calculations show that prioritizing waiters to achieve the optimal deviation from the RR rule may increase total sales by 9%. This gain primarily arises from an increase in the number of tables seated (enhanced operational productivity) instead of average sales per check (better service quality).

To the best of our knowledge, our paper is the first paper to empirically analyze how individuals not engaged in the work make routing decisions. Our results show that individuals adjust the routing rule that they are instructed to follow. Understanding the drivers of this deviation permits us to evaluate the routing decisions and identify opportunities for improvement not only for this setting, but also for better service systems in general. Finally, even though the router is often not someone with substantial training in that task (e.g., the host in our context), our results highlight the need to give attention and training to these front-line personnel (c.f. Tucker, 2007). Our paper answers the call by Boudreau et al. (2003) to incorporate the actions of people into the design of service systems - in other words to pursue a research agenda around people-centric operations. It is important to respond both to the needs of organizations today and in the future , particularly given the rise of rule/algorithm-based artificial intelligence (AI) in various service settings (Economist, 2016).

2 Related Literature

The customer seating problem in our paper is similar to routing problems in other areas of service operations. We first review the routing literature before discussing related research from the behavioral operations literature that examines how individuals deviate from predefined decision algorithms, optimal or not.

When constructing routing algorithms, a key question is the objective to optimize. Speed is an important measure of service operation performance in the routing literature. For example, in their seminal work, Lee and Cohen (1985) analyze how agents should direct customers to specific service facilities to optimize speed. However, there is often a trade-off between service speed and quality (Anand et al., 2011; Kostami and Rajagopalan, 2013). That is to say, it takes extra time to deliver higher service quality. This speed-quality trade-off is exacerbated by the heterogeneous servers who have different intrinsic skills in raising speed and quality, separately. Hence, researchers have created analytical models that design priority rules (e.g., De Véricourt and Zhou (2005); Armony (2005); Mehrotra et al. (2012); Zhan and Ward (2013)) to balance this speed/quality trade-offs.

Although these models may optimize the speed/quality trade-off, they can cause those servers having high skills to be more utilized than those servers having low skills (i.e., good work leads to more work). This imbalance in utilization may be perceived to be unfair towards servers. The perception of unfairness/injustice in the organization can cause various negative effects on worker performance (Cohen-Charash and Spector, 2001; Colquitt et al., 2001). Hence, fairness is another important consideration in designing job routing rules. The fairest routing rule is the round-robin rule because 1) it equally distributes work among servers, 2) does not discriminate task types and complexity, and 3) it prevents other strategic servers from cheating. Research finds that using the RR rule of assigning emergency department patients to physicians and nurses increases care providers' ownership of the patient and enhances their coordination, thus improving service quality and efficiency (Song et al., 2015; Valentine, 2018) Beside RR rules that maximize fairness, some analytical papers incorporate fairness into the objective function of the routing algorithm to balance the efficiency-fairness trade-off (Armony and Ward, 2010; Bertsimas et al., 2012; Geng et al., 2014; Mandelbaum et al., 2012; Ward and Armony, 2013).

These analytical models are insightful because they shed light on the difficult problem of routing work to heterogeneous servers while accounting for both the speed/quality trade-off and the efficiency/fairness trade-off. Optimal as these models/rules may be, they take for granted that decision makers will actually follow the algorithms after implementation. However, in many settings routing is not done automatically by a machine, but instead by a human. Humans are not algorithms; people are prone to their own incentives and heuristics. Even if an automated system is available, a decision maker may choose to sometimes bypass the system and make the decision manually (Van Donselaar et al., 2010). Some of these adjustments may hurt operational performance (Tucker, 2015), or alternatively, they may improve operational outcomes if frontline workers possess relevant information that is not accounted for in the implemented rule (Bowman, 1963). Hence, it is important to study empirically whether and how individuals actually adjust the implemented routing rule.

Our paper is also related to the literature that studies how humans deviate from predefined operational policies, optimal or not. This stream of literature tends to analyze supply chain management decisions, often in a controlled lab setting (e.g., Croson and Donohue, 2006; Kremer et al., 2011; Ovchinnikov et al., 2015; Schweitzer and Cachon, 2000; Sterman, 1989). Extending these experimental studies, our paper belongs to a growing stream of papers that start to use more observational data, which are typically difficult to collect, to study behavioral operations decisions. This type of work was originally called for by Bowman (1963) when he noted that by studying the choices that actual managers made, it was possible to, at times, find ways to improve performance. For example, Van Donselaar et al. (2010) study the ordering behavior of retail store managers and find that managers consistently revise automated order advice by advancing orders from peak to non-peak days. Store managers' revisions outperform the automated replenishment system because managers consider two factors ignored by the system: in-store handling costs and sales improvement potential through better in-stock. Similarly, Elmaghraby et al. (2015) analyze how salespeople in a B2B setting change the recommended decisions by considering cost-, product-, and customer-related factors that are not accounted for by the pricing tool recommendation. They do not consider the performance implications of the revision from the price recommendation, whereas we analyze the effect of deviation on restaurant performance outcomes. Cui et al. (2015) suggest that upstream decision makers may deviate from a given inventory policy, causing information losses in the ordering process. This insight explains the value of sharing downstream sales information in a consumer packaged goods company, where theoretical models otherwise predict no value of information sharing. Their paper does not explicitly model what makes the upstream decision makers deviate from the inventory policy, while we examine specific waiter-related factors that affect the deviation. All of these papers complement analytical models in assessing their assumptions, generalizability and implications. None of them has yet considered routing behavior in a service operations setting, which has generated many advanced analytical models. Our paper examines whether routers (restaurant hosts in our setting) deviate from an RR routing rule, which they are instructed to follow to ensure fairness and not to overload servers all at once. We conjecture that hosts consider three waiter-related factors to route incoming customers in the next section.

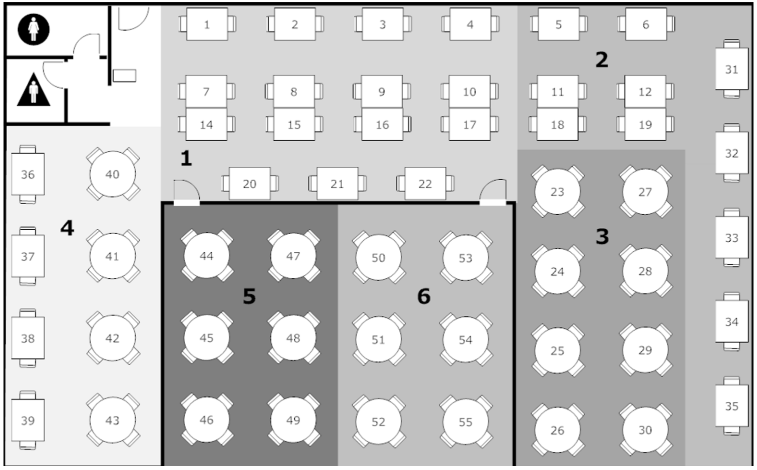
3 Hypothesis Development

3.1 The Role of Restaurant Hosts

Hosts greet the incoming customers and assign and walk them to a table. When the host finds no available tables or waiters, he/she will estimate a waiting time and manage a waiting list for incoming customers. According to our interviews with hosts, they tend to overstate the waiting time because they intend to set a low expectation about waiting for customers to avoid complaints. In short, the host is a gatekeeper. Other responsibilities of a host sometimes include wiping the front door glass and cleaning the toilet when the restaurant is not busy. When the restaurant is busy, some hosts volunteer to help waiters clear the tables, although clearing tables is the waiters' responsibility. The hosts are paid an hourly wage and do not share tips with waiters.

One primary responsibility of a host is assigning customers to different tables. The restaurant management instructs the hosts to follow the RR rotation rule (Walker, 2007, p. 229). That is, a host should direct incoming parties to each table section, which includes tables exclusively assigned to one waiter during a particular shift, by rotation like dealing a deck of cards. Figure 1 shows an example of the seating chart in a typical full service restaurant. A host follows the RR rule by rotating through each of the six table sections (labeled and shaded accordingly). This RR rule is deployed in the restaurants for three main reasons. First, it is easy to follow. Second, it should achieve strict fairness. Waiters' main income, tips, which is typically not shared with other waiters, depends on the number of and the types of parties that they receive. The RR rule equally distributes parties among them. In addition, it randomly assigns the types of party because the arrival process is random. The RR rule also prevents other waiters from strategically gaining or avoiding extra work. Third, the RR rule should pace and smooth the meal process. It avoids overloading a waiter by simultaneously seating multiple tables. Despite these advantages and its common use, we do not argue that the RR is either optimal to maximize total sales or generalizable in other service settings.

Figure 1: Restaurant Seating Chart



Our main research questions are 1) whether hosts deviate from the RR rule, 2) what waiter-related factors contribute to the improvisation, and 3) what is the implication of such adjustment for store performance? In this study, we do not intend to examine all factors that affect hosts' seating decisions, nor do we study how hosts put weights on these factors. As interesting as these questions are, we ask preliminary and yet important questions due to the lack of empirical studies on routing decisions. Focusing on elements about the waiter that are observed by the host, we conjecture three main factors, which are both theorized in the operations management literature and empirically testable with our data. These factors include waiters' contemporaneous workload, waiters' speed skills and waiters' sales skills. We seek to control for additional factors (e.g., the chemistry between the host and the waiter, the location of a table in the restaurant) that may bias our estimation of the three waiter-related factors in our main model and robustness checks.

3.2 Deviation from the RR Rule

To predict whether hosts deviate from the RR rule, we draw on the three theoretical dimensions of service improvisation competence (Secchi et al., 2019), which include creativity, spontaneity, and bricolage. Competent service workers may improvise in order to provide customers with satisfaction. In particular, they may “deviate from standard service delivery processes and routines” (creativity) to “respond to customers’ real-time unexpected requests” (spontaneity), “using available resources” (bricolage). In our restaurant setting, competent hosts should seat customers as quickly as possible to avoid or reduce customers’ waiting. They do so also to avoid manager’s reprimanding them for having too many customers waiting in the entrance area. For example, when a party randomly arrives and the waiter who is supposed to receive it is fully occupied, the customers may not want to wait for this busy waiter and want to be seated sooner. To respond to this request, the host may spontaneously assign them to another available waiter. In effect, the host creatively deviates from the standard RR rule. Admittedly, as the downside of deviating from the prescribed seating rule, waiters may directly complain to the host and/or to the manager about often being skipped. However, the immediate reward of seating the customers faster could outweigh the delayed cost of having an unpleasant conversation with the waiters or the manager because people tend to discount the later outcome or are present-biased (Thaler, 1981). Therefore, we posit:

H1: Hosts do not strictly follow the RR rule, and will improvise the seating rule.

3.3 Contemporaneous Workload

Even under an RR rule that guarantees fairness in terms of the total number of service requests assigned to each server, probabilistic customer arrival rates and uncertain service times can cause unbalanced contemporaneous workload (e.g., the number of patients that a doctor/nurse simultaneously takes care of or the number of tables that a waiter simultaneously handles) among servers. In this situation, the router may violate the RR rule to route the next service request to servers who have smaller contemporaneous workload for three reasons. First, allocating work to servers with the smallest workload, may reduce the total holding cost or waiting time of customers (e.g., Frostig and Levikson, 1999; Tezcan and Zhang, 2014). According to processor sharing theory (Parekh and Gallager, 1993), each of *n* jobs that a server simultaneously works on is likely to get approximately 1/*n*th of that server's limited capacity. In such a setting, a server having a small workload is therefore more capable of allocating more capacity to the newest task to expedite the service than a server under heavy load, controlling for everything else. In the restaurant setting, assigning an arriving party to a waiter with a current small workload is more likely to utilize that waiter's high remaining capacity to enhance service efficiency than would be the case of a waiter with a high workload.

Second, when workload is very high, servers may see their performance suffer (e.g., Bendoly and Prietula, 2008; KC and Terwiesch, 2009; Narayanan et al., 2009; Staats and Gino, 2012; Shah et al., 2016; Tan and Netessine, 2014) and so by allocating work to a server with lower workload, the quality of the completed work may be higher (e.g., in a restaurant, the customers may have a better experience). Similarly, seating an incoming party to a waiter under high workload, even strictly according to the RR rule, may overwhelm that waiter and spoil the dining experience for a customer.

Finally, uneven contemporaneous workload may be perceived as unfair or unequitable towards servers (e.g., Mandelbaum et al., 2012; Narasimhan et al., 2013). For example, high contemporaneous workload servers will have shorter idle (rest) time than low contemporaneous workload ones. Therefore, the router might allocate work to the lower contemporaneous workload server to avoid idling them and to make breaks more even. That is, the router may apply a work-conserving policy. Hosts may have such consideration, and consequently view allocating tables based on contemporaneous workload as a more fair and equitable choice, so as to spread the actual work out more evenly. For these reasons, we hypothesize:

H1: A host will direct more incoming parties to waiters having low workload than an RR rule suggests, all else being equal.

3.4 Speed Skills

Service speed is a key dimension upon which many service systems are evaluated. There are at least two primary reasons why servers with higher speed may be allocated more work. First, as discussed above, a router may seek to keep servers busy with an equal number of jobs. If a router were to use such a heuristic then servers that work faster would receive more tables over time. This can be thought of as analogous to

Little's Law (inventory = flow rate × flow time) where the number of jobs is analogous to inventory, flow time is analogous to service speed, and flow rate corresponds to jobs received from the router. Second, research suggests that routing service requests to the fastest servers first, whenever they are available, is optimal in various settings if the system objective is to reduce waiting time or to maximize service speed (e.g., Armony, 2005; Mandelbaum and Stolyar, 2004; Gurvich and Whitt, 2009). Thus, if a router seeks to improve performance on waiting time or service speed, the router may reallocate work to a faster server. In the restaurant setting, a host may give more tables to a prompt waiter in order to move customers through the restaurant quicker. Doing so can avoid customers' complaining about the long wait and managers' reprimanding them for many customers waiting in the entrance area. Therefore, we hypothesize:

H2: A host will direct more incoming parties to waiters with high speed skills than an RR rule suggests, all else being equal.

3.5 Sales Skills

Reducing waiting time or maximizing service speed is not the only objective that benefits a system. In many service settings, quality is not fixed, but instead a choice variable that a manager might seek to maximize. In order to accomplish this task, the service operations literature suggests that routing priority should be given to those servers having a combination of high speed and quality skills. For example, in order to minimize the average total time to resolve a call, De Véricourt and Zhou (2005) propose that the calls should be first routed to the servers having the highest effective service rate (i.e., resolution probability multiplied by service rate) whenever they are available. Note that this policy implies that priority should be given to quality conditioned on the same speed. Restaurants clearly want to maximize their sales (a quality measure) because the industry profit margin is between 3% and 9% (Walker, 2007). By allocating customers to individuals who sell more, holding speed constant, the restaurant can make more money. Although hosts have no access to sales data for each waiter, they see waiters carrying food and drinks. From this observation, they may infer waiters' sales ability. Hence, after we control for waiters' speed skills, we hypothesize:

H3: A host will direct more incoming parties to the waiters with high sales skills than an RR rule suggests, all else being equal.

3.6 Implications for Total Sales

H1 through H3 suggest that hosts may deviate from the RR rule, and assign more tables to waiters with lower contemporaneous workload, high speed and high sales skills. Such deviation could improve the total sales performance, the product of number of parties served and the average sales per party. For example, giving more parties to a waiter under low contemporaneous workload could utilize that waiter's high remaining capacity to enhance both service productivity and quality. In addition, prioritizing those waiters having high speed and sales skills could enhance both the throughput and the average sales per party.

However, these benefits of prioritization could have diminishing returns to the total sales. In fact, excessive deviation from the RR rule may eventually reduce the performance. First, the fewer waiters who receive most of the tables have their own capacity constraints, and so maybecome a bottleneck in the system. Continuously overloading such capacity-constrained waiters and underutilizing other available waiters may reduce the table turns, thus reducing the total sales. Second, giving a few waiters most of the tables may be perceived unequal and unfair, which may demotivate workers and reduce their service quality or speed (Cohen-Charash and Spector, 2001; Colquitt et al., 2001). Finally, excessive deviation from the RR unbalances the workload, causing either too much workload for some waiters or too little workload for the others. When the workload is either too high or two low, the sales performance deteriorates, according to Tan and Netessine (2014), who find an inverted-U shaped relationship between waiters' workload and sales performance. In other words, increased unequal table assignment will push the workload further down to the two lower suboptimal ends of the inverted-U performance curve. For these reasons, we hypothesize:

H4: As dispersion/inequality of table assignments increases (i.e., deviation from the RR), total sales first increase and then decrease: that is, there is an inverted-U-shaped relationship between the dispersion of table assignments and the total sales.

4 Empirical Setting, Data and Variables

4.1 Empirical Setting and Data

To test our hypotheses we use a large casual dining restaurant chain in a major U.S. metroplex as our empirical setting. We shadowed the hosts and conducted interviews with them, managers and waiters to gain institutional knowledge. When staffing the restaurants, the analytics department of the chain provides a sales forecast for the next week. It also applies a sales/waiter ratio to generate the forecast of the number of waiters needed each hour. After that, managers fill these shifts with available waiters in an online system. A lunch typically starts at 11 am and ends at 3 pm or 4pm. A dinner shift starts at 4 pm and ends at 8 pm. In addition, managers can schedule waiters or ask for volunteers to work closing hours (from 8 pm to restaurant closing generally at 11:00 pm) during the week. Unlike waiters, the demand for hosts is quite stable and only one host is staffed during a given hour.

In our setting, hosts are paid a fixed hourly wage, which does not include any sales performance components. In addition, waiters are paid a fixed hourly wage plus the tips from the customers they serve (typically 20% of the sales amount). They do not share the tips with the host, alleviating the potential omitted variable concern of prioritizing certain waiters out of financial incentives. However, hosts may occasionally receive a tip from the customers, probably for them to get seated sooner. Although this incentive changes which customer party may get prioritized, it does not affect which waiter receives tables earlier than the RR rule, the topic of this paper. Managers are paid a salary and a store performance incentive. Therefore, understanding how to train hosts and improve sales performance is an important managerial issue.

In our interviews, we saw the host implement the RR rule with a dining room seating chart of color-coded numbered table sections on a screen at his/her station (similar to Figure 1). The color coding and numbering helps the hosts remember the seating sequence, such as blue-orange-red-pink or 1-2-3-4. The screen also shows the current number of tables seated in each table section, and pops up the next table section to seat according to the RR rule. Note that our focal restaurant chain does not take reservations, which makes our setting relatively "pure" to study hosts' possible deviation from the RR rule. In addition, the table sections in the dining room should be equally amenable to various parties of customers because each section has a mixture of booths and tabletops, four-top tables and two-top tables.

How is the RR rule enforced? Waiters sometimes self-police the table assignment and will complain to the hosts or escalate to the managers if they feel the assignment is not done fairly (i.e., the RR rule is frequently violated). Managers may talk with the host, and even reduce his/her shifts if that host receives repeated complaints. Despite these penalties and the fact the hosts want to be fair, they can exercise their discretion about where to seat the incoming customers. Customers sometimes also influence the hosts to change their seating assignment because of table or waiter preference. Hence, it prompts us to conduct a rigorous study to understand how hosts make such discretionary seating decisions and how those decisions impact performance.

We gathered point-of-sales (POS) data and labor working time data of nine-month length. We received the data in three discontinuous brackets of months because of data confidentiality concerns on the part of the company. The data include detailed information about hosts, waiters, sales, party size, and service start and end time for each check. We combined the split checks within a party, so one table has only one check. Since the hosts in our restaurants neither handle take-out orders nor seat customers to particular bartenders in the bar, we focus only on the dining room data. Our data set is comprised of approximately 1,964,489 check-level observations. Ideally, we would like to observe waiting time data, but we could not obtain that since current restaurant technology usually does not record the waiting time information. We also wish we possessed the electronic records of the RR sequence and the actual seating time. Due to these data restrictions, we choose to use hour-level aggregation of the number of checks opened to infer whether the host equally assigns tables to waiters according to the rotation rule (more details are provided in subsection 4.2.1).

Our casual dining data provide a valuable empirical setting to study whether and how a router deviates from a rule that he/she is instructed to follow and to evaluate such discretionary deviation. First, waiters are a heterogeneous group, with varying sales and speed skills (Tan and Netessine, 2015). These variations allow us to understand how server-related factors may contribute to a router's discretionary assignment preferences. Second, unlike call centers where callers may have various priority levels and complicated needs (e.g., making reservations, filing complaints, asking technical questions), casual dining customers generally receive equal treatment and have relatively simple needs (i.e., to eat and drink), allowing our study to focus on the impact of heterogeneous waiters rather than customers on hosts' discretionary routing decisions. In other words, our empirical setting reflects routing a single type of job to heterogeneous servers. Admittedly, the varying party size of each party complicates the seating tasks. We attempt to control for such factors in our following empirical analysis. Finally, the waiter assignment process in the restaurants is common to a wide variety of task routing settings, where customer arrival within a fixed time period is random and servers are heterogeneous. This similarity not only increases the relevance of our study to widely studied analytical models, but also makes it possible to generalize our managerial implications.

4.2 Variables

4.2.1 Dependent Variable

$HrTableDiff_{jt}$. The goal of our analysis is to examine whether hosts equally assign incoming parties to waiters according to the the RR routing rule. To operationalize this measure, we compute the difference between the number of tables that waiter *j* is assigned and the average number of tables per waiter during the hour *t*. If this gap is positive (negative), that waiter should receive more (fewer) tables than the RR rule suggests.

Ideally, we would like to analyze check-level data because this level of granularity can typically reveal more nuance about hosts' routing behavior. Nevertheless, we do not have the data about customers' arrival time, waiting time, and seating time. We observe when a check is opened by the waiter instead, which does not always coincide with the actual seating time. Approximately 20% of the time (according to an interviewee), waiters may open the checks in batches, especially when the restaurant is busy and customers keep getting seated. In addition, some later seated customers may be ready to order earlier than those who arrived earlier. Some waiters may also open the checks more promptly than other waiters. All these factors may cause measurement error if we use check opening time to infer the check-level seating sequence.

To address this important issue, we use hour level aggregation as our main analysis. First, the typical delay between customers' getting seated and waiters opening the check should be on average quite short (perhaps less than ten minutes based on our interviews). In fact, check opening time is a common measure to infer the beginning of a meal service in the restaurant operations literature (e.g., Kimes, 2004). Therefore, one-hour aggregation becomes less sensitive to such measurement error. Furthermore, if the delays are similar across the waiters, check opening time should still reflect the seating sequence. The dispersion of such delays may cause measurement error bias; yet, this dispersion of delays is even shorter than the average delay. Hence, the hour-level aggregation becomes even less sensitive to such measurement error. Finally, the aggregate data should average out such heterogeneous delay differences, making our hour-level measures unbiased. In addition to using hourly aggregated data, we include waiter fixed effects in our models to control for a waiter's innate tendency towards check opening delay. We also conduct robustness checks using a waiters' table difference divided by the hourly average (a scale-free relative measure) and check-level observations. Together these checks show the same pattern of results and provide assurance that our dependent variable should represent the direction and the size of the potential deviation from the rotation rule.

4.2.2 Independent Variables

$TableLoad_{jt}$. We follow KC and Terwiesch's (2011) method to compute waiter *j*'s average workload during hour *t*. In particular, we first calculate the number of tables (parties) that waiter *j* simultaneously handles at the beginning and at the end of hour *t*, respectively. Then, we take the average of these two. We used the number of tables at the beginning of the hour as a robustness check and the results are consistent.

$SpeedSkill_{jm}$. Following Mas and Moretti (2009) and Chan et al. (2014), we employ a fixed-effects model to estimate the intrinsic speed skill of the waiter *j* in month *m*. The speed skill level is assumed to be constant within the same month but changing over time for reasons such as learning (e.g., Argote and Epple, 1990), forgetting (e.g., Shafer et al., 2001), and task variation (e.g., Staats and Gino, 2012). In particular, we first divide our data into nine months. We then specify the following fixed effects model to estimate $SpeedSkill_{jm}$, the speed ability of waiter *j,* and repeat this model estimation for each month *m,* separately:

$$\begin{aligned} \mathit{AvgMealDuration}_{jtm} &= \alpha_0 + \mathit{SpeedSkill}_{jm} + \alpha_1 \mathit{AvgPartySize}_{jtm} + \alpha_2 \mathit{HrTables}_{jtm} + \\ & \alpha_3 \mathit{PPA}_{jtm} + \alpha_4 \mathit{Controls}_{tm} + \xi_{jtm} \quad \forall m = 1, \cdots, 9. \end{aligned} \tag{1}$$

In this model, $\mathit{AvgMealDuration}_{jtm}$ is the average meal duration (in minutes) of all the checks opened by waiter j during hour t in month m. Meal duration has been approximated in the literature by the difference between check opening and closing times recorded in POS data (e.g., Kimes, 2004; Tan and Netessine, 2014). A large value of $\mathit{SpeedSkill}_{jm}$ indicates that the waiter is actually *slow* because *SpeedSkill* is essentially a fixed effect of additional time added to (or subtracted from) meal duration. In addition, we control for the average party size of these checks (*AvgPartySize*) and workload (*HrTables,* the number of tables/parties assigned to waiter j during hour t). These factors are shown to affect meal duration (Tan and Netessine, 2014). Furthermore, we compute PPA_{jtm}, the per person average dollar sales in each hour by averaging the sales of all the checks opened by waiter j during the hour over all the customers who contribute to these sales. We control for $\mathit{PPA}_{jtm},$ because sales can affect meal duration (e.g., customers may take extra time to eat a larger and often more expensive meal). Conditioning on the sales makes the additional variation in meal duration exclusively attributed to a server's promptness. Admittedly, the coefficient of PPA_{jtm} is subject to endogeneity bias because meal duration can also affect PPA_{jtm}. Nevertheless, identifying the causal effect of sales is not our goal. Rather, we solely use PPA_{jtm} as a control variable for the speed-meal duration trade-off. Furthermore, we use $\mathit{Controls}_{tm}$, including $\mathit{DayWeek}_{tm},$ $\mathit{Hour}_{tm},$ Trend_{tm} and $\mathit{Store}_{tm},$ to adjust for temporal and locational factors about meal duration. In particular, we include a categorical variable *Hour,* the hour of the day, to control for time-related factors, such as demand within the day. Categorical control, *DayWeek* indicates the day of the week since weekends are usually busier than weekdays. Linear continuous variable *Trend* (calendar day) adjusts for daily trends. Categorical variable *Store* controls for the time-invariant aspects of store fixed effects (e.g., location, traffic).

$\mathit{SalesSkill}_{jm}$. Similar to estimating waiters' speed skills, we employ the following fixed-effects model to estimate the sales premium, or intrinsic sales skills of waiter *j* during hour *t*:

$$\begin{aligned} PPA_{jtm} =& \beta_0 + SalesSkill_{jm} + \beta_1 AvgPartySize_{jtm} + \beta_2 HrTables_{jtm} + \\ & \beta_3 AvgMealDuration_{jtm} + \beta_4 Controls_{tm} + \varepsilon_{jtm} \quad \forall m = 1, \cdots, 9. \end{aligned} \tag{2}$$

In this model, a large value of $SalesSkill_{jm}$ means that the waiter has a high sales ability. We control for $AvgMealDuration_{jtm}$ because staying at the restaurant longer creates more opportunities for customers to order more food/drink items. Conditioned on the meal duration, the additional variation in sales should be attributed to a waiter's sales ability, or "sales productivity". Similar to Model 1, we further include workload $HrTables_{jtm}$ and $Controls_{tm}$, and repeat the model estimation every month to account for learning, forgetting and task variation.

We conduct five additional robustness checks to alternatively estimate the two skill variables. These alternative calculations include: 1) assessing the entire nine months together, 2) evaluating the skills every week, 3) removing *PPA* and *AvgMealDuration* as control variables from Models 1 and 2, respectively, 4) controlling for staffing in both models, and 5) using random effects models for estimation. All the results are shown in the online Appendix, and are consistent with the current skill definitions.

4.2.3 Descriptive Statistics

Table 1 shows the descriptive statistics of the data. The restaurants in our data set are comparable to other casual dining restaurants. An average party has 2.01 people. They spend on average \$31.10 and 52 minutes on the meal. There are 4.35 waiters working every hour, each of whom is assigned 2.24 tables on average. Nevertheless, some waiters receive as many (few) as three tables more (less) than the average. Furthermore, the waiters have heterogeneous speed and sales skills ranging from -12.79 minutes to 19.18 minutes, and from -\$2.15 to \$2.26, respectively.

Table 2 shows the comprehensive correlation matrix. *HrTableDiff* is positively associated with *TableLoad* (correlation = 0.3295) and *SalesSkill* (correlation = 0.0167), and negatively correlated with *SpeedSkill* (correlation = -0.0433). In addition, *SpeedSkill* and *SalesSkill* are negatively correlated (correlation = -0.2475),

Table 1: Summary Statistics

	Sales	*MealDuration*	*PartySize*	*HrWaiters*	*HrTableDiff*	*TableLoad*	*SpeedSkill*	*SalesSkill*
Mean	31.10	52.00	2.01	4.35	-.00	2.24	-0.23	0.00
Stdev	15.50	19.83	1.03	1.92	1.08	1.36	6.50	0.91
Min	6	21	1	1	-3	0	-12.79	-2.15
P50	27.89	48	2	4	0	2	-0.87	-0.02
Max	200.72	131	6	10	3	6	19.18	2.26

suggesting that a fast waiter may also have high sales productivity, conditioned on meal duration (all correlations, p<0.05). Notably, the correlation between *TableLoad* and *SalesSkill* is insignificant and close to zero, and the correlation between *TableLoad* and *SpeedSkill* is 0.0806, very small, too. The low correlations are due to our controlling workload (*HrTable*s) when estimating both *SalesSkill* and *SpeedSkill* in Models 1 and 2, which alleviates the potential multicollinearity in these variables. In general, the correlations among the independent variables are all significantly below 0.8, which provides some assurance that they should not cause multicollinearity. These correlations do not control for other factors of waiter assignments. Thus, we turn to Section 5 for a more rigorous empirical analysis.

We additionally report the correlations of continuous daily trend (*Trend*), the number of waiters per hour (*HrWaiters*), the total number of tables that opened checks during the hour (*TotalTables*), the average hourly workload in terms of tables (i.e., $AvgTableLoad = TotalTables/HrWaiters$), and the average party size during the hour. None of these correlations are either greater than 0.8, or less than -0.8. Although the correlation between *TotalTables* and *HrWaiters* (0.7456) and the correlation between *TotalTables* and *AvgTableLoad* (0.6506) are still within the bounds of 0.8 and -0.8, they are moderately high because restaurants use traffic forecasts to staff waiters. We therefore use only *TotalTables* as the moderator of store demand without including either *HrWaiters* or *AvgTableLoad* in the moderating effect analysis (Model 8).

5 Empirical Strategy and Results

5.1 Main Effects

To answer the first question whether hosts deviate from the RR rule, we plot the histogram of *HrTableDiff*. If there is a significant dispersion, then we will conclude that hosts do not always adhere to the RR rule. Then,

Table 2: Correlation Matrix

	HrTableDiff	*TableLoad*	*SalesSkill*	*SpeedSkill*	*Trend*	*HrWaiters*	*TotalTables*	*AvgTableLoad*
TableLoad	0.3295*	1.0000						
SalesSkill	0.0167*	0.0000	1.0000					
SpeedSkill	-0.0433*	0.0806*	-0.2475*	1.0000				
Trend	-0.0000	-0.0263*	-0.0001	0.0002	1.0000			
HrWaiters	0.0011	0.0154*	0.0212*	-0.0375*	-0.1498*	1.0000		
TotalTables	0.0009	0.2042*	0.0013	-0.0600*	-0.0813*	0.7456*	1.0000	
AvgTableLoad	0.0003	0.3121*	-0.0182*	-0.0508*	0.0544*	0.0452*	0.6506*	1.0000
AvgPartySize	0.0003	-0.0043*	-0.0158*	-0.0244*	0.0868*	0.1195*	0.0411*	-0.0639*

*Significant at the 0.05 level

we employ the following model to examine whetherthe three proposed waiter-specific factors contribute to the deviation:

$$\begin{aligned} HrTableDiff_{jt} &= \alpha_0 + \alpha_1 TableLoad_{jt} + \alpha_2 SpeedSkill_{jm} + \alpha_3 SalesSkill_{jm} + \\ &\quad \alpha_4 WaiterFE_j + \alpha_5 HostFE_j + \alpha_6 Controls_{tm} + \varepsilon_{jt} \qquad \forall m = 1, \cdots, 9. \end{aligned} \tag{3}$$

If the coefficient of any of the three independent variables is significantly different from zero, then that independent variable may contribute to the deviation. For example, a negative α_1 suggests that the host may give more tables to a waiter under lower contemporaneous load. $WaiterFE_j$ and $HostFE_j$ are waiter and host fixed effects, respectively. They control for unobserved waiter and host heterogeneity, such as the aforementioned waiter-specific check opening delays, waiters' idiosyncratic popularity among customers and hosts or the probability of being requested. $Controls_{tm}$ represents the same control variables as in Model 1, which adjust for temporal and locational factors. We also cluster standard errors in terms of host and hourly sequence to allow for correlation within hosts' assignments and heteroskedastic errors over time.

In this model, giving a waiter disproportionately more tables than average will inflate this waiter's workload, thus creating an upward bias in estimating α_1. In addition, some omitted variables including waiters' good personality, attractiveness or athleticism may be correlated with the speed and sales skills and the hosts' waiter assignment preference, causing an omitted variable bias. In order to alleviate such endogeneity concerns, we control for waiter and host fixed effects, and employ an instrumental variable two-stage-least-square (2SLS) approach, which can provide consistent estimates of the dependent variables with a large sample (Angrist and Krueger, 1994).

A valid instrumental variable should satisfy both the relevance and exclusion restriction conditions (Wooldridge, 2002). The relevance condition means that the instrument should be correlated with the endogenous variable (e.g., *TableLoad*). The exclusion restriction condition requires the instrument to affect the dependent variable (i.e., *HrTableDiff*) only through the endogenous variable. We introduce a type of instrumental variable, which should satisfy these two conditions.

Similar to prior work (e.g., Bloom and Van Reenen, 2007; Siebert and Zubanov, 2010), we use the lagged value of the endogenous independent variable as an instrument. For $TableLoad_{jt}$, we follow Tan and Netessine's (2014) approach and compute the average storewide hourly workload during the same hour as t (i.e., the total number of checks opened divided by the number of waiters). Then, we compute the one week lagged value of this workload. For example, for a check that is opened at 12:30 pm on 1/8/2013, its instrument is the average hourly workload of the same store at the 12:00 pm slot on 1/1/2013. For $SpeedSkill_{jm}$ and $SalesSkill_{jm}$, we cannot compute one week lagged value as instrumental variables because these variables are constant within the same month. Instead, we use one month lagged value of waiter j's speed and sales skills.

To implement these three instruments, we first use them to estimate the 2SLS, separately. Then, we include all of these instruments together to show consistency of our results. These lagged variables should satisfy the relevance condition because 1) the coefficient during the first stage estimation turns out to be significant and positive, suggesting a positive correlation with the current week or month in the first stage estimation, and 2) the first-stage F-statistics turns out to considerably exceed 10, the rule of thumb for weak instruments in every model (Staiger and Stock, 1997). Furthermore, these instruments should satisfy the exclusion restriction condition because the storewide workload one week prior or the waiter's speed and sales skills a month prior should not affect the host's table seating behavior the current week or the current month, after controlling for the temporal factors.

5.2 Results

Figure 2 shows the histogram of the gap between the number of tables assigned to each waiter and the hourly average in our data (i.e., *HrTableDiff*). This histogram suggests a wide variance in the number of tables received by each waiter each hour (standard deviation is equal to 1.08). In particular, if we consider *HrTableDiff* = 0 as strictly following the RR rule, hosts are found to deviate approximately 90% of the time. If we define $|HrTableDiff| \leq 1$ as adherence, hosts improvise approximately 28% of the time. We also conduct a Kolmogorov-Smirnov equality-of-distributions test to test whether *HrTableDiff* follows a uniform distribution between -1 and 1. The *p*-value turns out to be less than 0.001, which rejects the hypothesis. Hence, we find support for H1 that predicts hosts do not strictly follow the RR rule.

Figure 2: Histogram of the Gap between Employee's Hourly Assigned Table Quantity and the Hourly Average

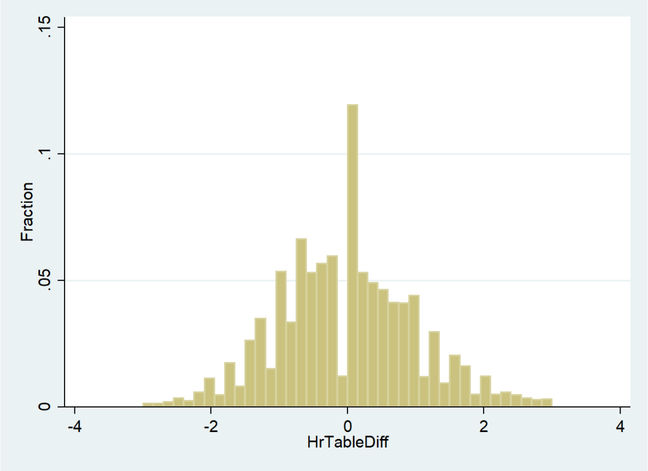


Table 3 shows the results of our hour-level analysis of whether the three proposed waiter-specific factors contribute to the hosts' deviation from the RR rule. The coefficient of *TableLoad* is positive in the models that treat it as exogenous (0.3451 in column 1, 0.3473 in column 3, and 0.3472 in column 4). However, the positive coefficient is corrected by the 2SLS in the expected direction and becomes negative in the models that treat it as endogenous (-0.0622 in column 2, and -0.0518 in column 5). The negative sign suggests that a host may give fewer tables than average to a waiter under heavy load, supporting Hypothesis 2. Interpreting the coefficient in column 2, having one standard deviation (1.36) additional table of workload makes a host give that waiter $1.36 \times 0.0622 \approx 0.085$ tables below the average. This effect size is also organizationally significant because it is approximately 10% of the mean absolute value of *HrTableDiff* (i.e., $E(|HrTableDiff|)$ =0.8 tables/hour).

In addition, the coefficients of *SpeedSkill* are consistently significant and negative (the coefficients are -0.0141, -0.0053, -0.0147, -0.0139 and -0.0045). The significant and negative signs suggest that hosts may give faster waiters more tables than average, consistent with Hypothesis 3. However, high sales skills waiters may not receive more tables than average. Controlling for everything else, the coefficients of *SalesSkill* are not statistically significant in all models, which fails to support Hypothesis 4. Interpreting the 2SLS coefficients for *SpeedSkill* in column 3, a waiter who has speed skills one standard deviation (6.64 minutes) higher may be assigned $6.64 \times 0.0147 \approx 0.098$ tables above the average. Similar to the effect size of *TableLoad,* this difference size is considerable as it is $0.098/0.8$ =12.25% of the mean absolute value of *HrTableDiff.* In this analysis, we observe that hosts assign faster waiters, but not necessarily high sales ability ones, more tables. This is probably because hosts have no access to POS data to accurately infer waiter's sales productivity, but can observe who works promptly. In other words, waiters' speed skills are more observable to the hosts than their sales skills. In addition, hosts do not have as strong an incentive to increase sales as to seat customers as swiftly as possible. Nevertheless, we do not conclude that hosts absolutely ignore a waiter's sales ability because drawing that conclusion may require an even larger sample. Instead, we argue that hosts are more likely to consider waiters' speed skills than their sales skills when assigning tables. As a robustness check, we include the three independent variables separately to avoid multicollinearity. The results are presented in the online Appendix, and are congruent with the results of including all the variables together as shown here.

Finally, the results in column 5 are consistent with those in column 2. We argue that *SpeedSkill* and *SalesSkill* are less likely than *TableLoad* to be subject to endogeneity bias. These skills are conditional estimates given workload ($HrTables_{jtm}$ in Models 1 and 2), which account for whether waiters receive tables equally. In addition, unlike reverse causality, a host's table assignment does not change waiters' intrinsic speed and sales skills during hour *t* because these waiters' long-term characteristics are time-invariant characteristics such as gender and height. Admittedly, there may be omitted variables that are related to both the skills and the host's table assignment decisions. These variables are more likely to be waiter- or host-specific factors, such as the waiter popularity, the chemistry between the host and the waiter or favoritism. They are already controlled for through their fixed effects in the models. For these reasons, we choose to treat only *TableLoad* as an endogenous variable in the rest of the analysis.

Table 3: Hour-level Table Assignment

	Column 1: Model 3 Estimated by OLS	Column 2: Model 3 Estimated by 2SLS	Column 3: Model 3 Estimated by 2SLS	Column 4: Model 3 Estimated by 2SLS	Column 5: Model 3 Estimated by 2SLS
TableLoad	0.3451***	-0.0622***	0.3473***	0.3472***	-0.0518***
	(0.0042)	(0.0064)	(0.0050)	(0.0050)	(0.0090)
SpeedSkill	-0.0141***	-0.0053***	-0.0147***	-0.0139***	-0.0045*
	(0.0005)	(0.0006)	(0.0014)	(0.0011)	(0.0022)
SalesSkill	0.0026	0.0021	0.0021	0.0090	0.0072
	(0.0019)	(0.0026)	(0.0027)	(0.0233)	(0.0312)
WaiterFE	Yes	Yes	Yes	Yes	Yes
HostFE	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
Observations	705,575	595,189	436,063	436,063	397,114
Prob>Chi-sq	<.001	<.001	<.001	<.001	<.001

1. Clustered standard errors at host and hour sequence level are shown in parentheses. 2. *$p\leq$ .05, **$p\leq$ .01, ***$p\leq$ .001. 3. In Columns 2 through 4, *TableLoad, SpeedSkill* and *SalesSkill* are treated as endogenous and instrumented, respectively. In Columns 5, these variables are instrumented altogether.

6 Robustness Checks and Post-hoc Analysis

To gain confidence in the results presented in Table 3, we perform the following robustness checks that examine alternative dependent variables. Below, we describe each model before reporting the results in Table 4.

Percentage Table Assignment Differences The average number of tables assigned to waiters fluctuates every hour, which may affect the sizes of table assignment differences. To have a scale-free deviation measure, we use percentage table assignment differences as an alternative dependent variable. In particular, we define *PctHrTableDiff* as the ratio between *HrTableDiff* and the hourly average number of tables assigned. We then use the same independent variables and control variables to reestimate Model 3. The results of using this alternative dependent variable is shown in Column 1 of Table 4, and are consistent with the main results.

Check-level and Half-hour-level Alternative Dependent Variables Granular-level analysis can typically adjust for more nuanced information. In our case, some customers may prefer or dislike certain tables, and persuade the host to change their table assignment. In our data, we observe a unique ID for each table. To control for the table fixed effects, we conduct a check-level analysis. In this analysis, we construct a new dependent variable called $RRGap_i$. We first order all the checks during a shift (lunch or dinner) at a restaurant by check open time (a proxy for the storewide seating sequence) and assign them a sequence number. For each check *i*, we then compute the theoretical order of the next assignment of this waiter according to the RR rule, which equals the current sequence number plus the number of waiters during this hour. For example, for the 15th check of a shift, when eight waiters work together during that hour, the next theoretical RR assignment order of this waiter should be the (15+8=23rd) check of the shift. Then the difference between the theoretical order and the actual order is defined as *RRGap.* A positive *RRGap* implies that a host will assign that waiter earlier than RR rule suggests (i.e., more tables for this waiter). Furthermore, we replace independent variable $TableLoad_{jt}$, the average number of tables waiter *j* handles during hour *t,* with $ContempTables_i$, the contemporaneous number of tables that the waiter simultaneously handles before assigned to party *i.* We include table fixed effects and keep the other independent variables and control variables the same. That is,

$$\begin{aligned} RRGap_i = &\alpha_0 + \alpha_1 ContempTables_i + \alpha_2 SpeedSkill_i + \alpha_3 SalesSkill_i + \\ &\alpha_4 WaiterFE_i + \alpha_5 HostFE_i + \alpha_6 TableFE_i + \alpha_7 Controls_i + \varepsilon_i. \end{aligned} \tag{4}$$

Admittedly, the check-level dependent variable *RRGap* assumes that the check opening sequence is a proxy for table assignment sequence. This assumption will be questioned, for example, when waiters open the checks in batches, especially when the restaurant is busy and customers keep getting seated. Our main hour-level analysis in Model 3 relaxes this assumption. First, the hour-level aggregation is less sensitive to the heterogeneous delays of opening the checks among waiters. Second, the aggregate data should average out such heterogeneous delay differences. For these reasons, we elect to use the hour-level analysis as our main analysis and the check-level analysis as a robustness check.

To find a middle ground between check-level and hour-level analyses, we alternatively examine the half-hour-level differences of table assignments. We define the dependent variable $\textit{HalfHrTableDiff}_{jh}$ as the difference between the number of tables that waiter j is assigned and the average number of tables per waiter during the half hour h. Accordingly, we change the independent variable to $\textit{HalfHrTableLoad}_{jh}$, which represents the waiter j's average workload during the half hour h. That is, the average between the numbers of tables (parties) that waiter j simultaneously handles at the beginning and at the end of half hour h, respectively.

Columns 2 and 3 in Table 4 show the results of check-level and half-hour-level alternative dependent variables. As can be seen, the results are qualitatively congruent with the hour-level main results.

Hour-level Alternative Dependent Variable in Terms of Customers A natural alternative dependent variable is the difference in terms of the customer quantities assigned to the waiters. We define $\textit{HrCustomerDiff}_{jt}$ as the difference between the number of customers that waiter j is assigned and the average number of customers per waiter during the hour t. We keep the independent variables and the control variables the same as in Model 3. Column 4 of Table 4 presents the results of hour-level alternative dependent variable in terms of customers. The results also show consistent patterns to the main results.

Post-hoc Analysis: Moderating Effects of Store Demand and Party Size High traffic intensifies the arrival uncertainty and meal duration uncertainty, making it difficult to follow the RR rule. For example, during non-busy hours (total hourly number of checks opened is less than the mean of 12) hosts follow the RR rule 82%[2] of the time, but during busy hours (greater than 12 checks opened) hosts follow the RR rule only 66% of the time. Furthermore, restaurants perceive large parties as more challenging than small parties because there are additional service requests from large parties. Hosts may turn to "stronger" or "capable" waiters to accommodate these customers. For these reasons, we employ the following models to understand the moderating effects of the storewide traffic and the size of incoming parties on host's waiter assignment

[2] We assume $|\textit{HrTableDiff}|$ to be adherence.

Table 4: Robustness Checks

	Column 1:	Column 2:	Column 3:	Column 4:
	Model 3 with DV	Model 4 with DV	Model 3 with DV	Model 3 with DV
	= *PctHrTableDiff*	= *RRGap*	=	=
			HalfHrTableDiff	*HrCustomerDiff*
TableLoad	-0.0043**			-0.1589***
	(0.0014)			(0.0430)
ContempTables		-0.3853***		
		(0.0440)		
HalfHrTableLoad			-0.0256***	
			(0.0005)	
SpeedSkill	-0.0003***	-0.0047**	-0.0012***	-0.0140***
	(0.0001)	(0.0018)	(0.0002)	(0.0015)
SalesSkill	-0.0001	-0.0004	0.0010	-0.0996
	(0.0002)	(0.0052)	(0.0010)	(0.0733)
Host FE	Yes	Yes	Yes	Yes
Waiter FE	Yes	Yes	Yes	Yes
TableFE		Yes		
Controls	Yes	Yes	Yes	Yes
Observations	595,189	1,509,299	1,134,248	595,189
Prob>Chi-sq	<.001	<.001	<.001	<.001

1. Standard errors are shown in parentheses. 2. *$p\leq$ .05, **$p\leq$ .01, ***$p\leq$ .001.

behavior:

$$\begin{aligned} HrTableDiff_{jt} =& \alpha_0 + \alpha_1 TableLoad_{jt} + \beta_1 TotalTables_t + \mu_1 TableLoad_{jt} \cdot TotalTables_t + \alpha_2 SpeedSkill_{jm} + \\ & \mu_2 SpeedSkill_{jm} \cdot TotalTables_t + \alpha_3 SalesSkill_{jm} + \mu_3 SalesSkill_{jm} \cdot TotalTables_t + \\ & \alpha_4 WaiterFE_j + \alpha_5 HostFE_j + \alpha_6 Controls_{tm} + \varepsilon_{jt}, \end{aligned} \tag{5}$$

$$\begin{aligned} HrTableDiff_{jt} =& \gamma_0 + \gamma_1 TableLoad_{jt} + \theta_1 AvgPartySize_{jt} + \nu_1 TableLoad_{jt} \cdot AvgPartySize_{jt} + \gamma_2 SpeedSkill_{jm} + \\ & \nu_2 SpeedSkill_{jm} \cdot AvgPartySize_{jt} + \gamma_3 SalesSkill_{jm} + \nu_3 SalesSkill_{jm} \cdot AvgPartySize_{jt} + \\ & \gamma_4 WaiterFE_j + \gamma_5 HostFE_j + \gamma_6 Controls_{tm} + \xi_{jt} \qquad \forall m = 1, \cdots, 9. \end{aligned} \tag{6}$$

The storewide traffic is defined as the total number of tables that opened checks during hour *t*, $TotalTables_t$. Ideally, we would like to include the number of waiting customers; however, we do not have this information. In addition, we define $AvgPartySize_{jt}$ as the average number of customers in a party that was assigned to waiter *j* during hour *t*. The rest of the control variables are identical to the control variables in Model 3.

Table 5 presents the moderating effects of store demand and party size. In the store demand moderating effect model (column 1, Model 5), the coefficient of *TableLoad* is significant and negative (-0.1015), consistent with the primary results in Table 3. In addition, the coefficient of *TotalTables* (mean is 14.49; standard deviation is 6.89) is significant. The lack of significance is expected because $HrTableDiff_{jt}$ should cancel each other out within an hour, when *TotalTables* is constant for all the waiters. The interaction term with *TableLoad* is also not statistically significant (coefficient = 0.0003), suggesting that store demand does not seem to moderate the effect of contemporaneous workload on the dispersion of table assignments. Furthermore, the coefficient of *SpeedSkill* is significant and positive in Model 5 (0.003), while the coefficient of its interaction with *TotalTable* is significant and negative (-0.0005). These coefficients first suggest that under increasing store demand, a host becomes more likely to direct these customers to a faster waiter. In addition, under low demand (when *TotalTables* is below six tables), a host may give more tables to a slow waiter, possibly because 1) the host intends to compensate the slower waiter with additional tables that he/she missed during busy hours to be fair, and 2) the host may think that the slow waiter can better handle the tables during slow hours. We note that our results in our primary analysis dominate as hosts largely give more tables to faster waiters on average because the tipping point (six tables) is approximately $(14.49-6)/6.89 \approx 1.23$ standard deviations below the sample mean of *TotalTables*, which constitutes only approximately 10% of the observations. In addition, the coefficients of *SalesSkill* and the interaction term with *TotalTables* are not significant, which suggests that a host does not seem to take a waiter's sales skill into his/her table assignment decisions facing varying store demand.

In the party size moderating effect model (column 2, Model 6), the coefficient of *TableLoad* is significant and equal to 0.3433, while the coefficient of its interaction with *AvgPartySize* (mean is 2.04, standard deviation is 0.75) is significant and negative (-0.2052). These results imply that hosts tend to give more large incoming parties (greater than 0.3433/0.02052$\approx$ 1.7 customers per party, representing approximately 75% of the data) to waiters having low contemporaneous workload. These waiters are more likely to have capacity to handle large parties, which typically require more work. These waiters are also more likely to have empty tables in their sections to combine and accommodate big parties. On the flip side, when the party size is small (fewer than 1.7 customers), the host may give more tables to waiters having high contemporaneous workload to compensate them and to be fair. Since the large incoming parties (greater than 1.7 customers) compose the majority of the observations, our primary result suggests that the host generally gives more tables to waiters having low workload on average. Furthermore, the coefficient of *SpeedSkill* is significant and negative (-0.0149), consistent with the primary result in Table 3. The coefficient of the interaction term with *AvgPartySize*, however, turns out to be significant and positive (0.0047), suggesting that a high speed ability waiter may receive a table later when his/her party size is very large (greater than 0.0149/0.0047$\approx$ 3.17 customers per party, representing less than 10% of the data). It is possible that hosts are intentionally assigning large and small parties differentially based on speed. It may be that hosts try to ensure a sense of fairness in terms of the total number of customers because fast waiters tend to receive more tables on average. It may also be that large parties require more attention and have less variation in the meal duration. Indeed, the coefficient of variation of the meal duration for party size above the mean is 0.34, while the coefficient of variation of the meal duration for party size below the mean is 0.39. This would suggest that the host may either prioritize giving faster workers small parties or slower workers large parties. We also note that the tipping point for party size of 3.17 is approximately $(3.17-2.04)/0.75 \approx 1.5$ standard deviations above the sample average party size and parties having over 3.17 customers constitute less than 10% of the observations; so, by and large, we see that hosts tend to give more tables to faster waiters. Finally, the coefficient of *SalesSkill* is significant and equal to -0.0774, while its interaction term is significant and equal to 0.037. These results suggest that a high sales ability waiter may receive fewer (more) tables when his/her average party size is less (greater) than 0.0774/0.037$\approx$ 2.09. Intuitively speaking, it is plausible that hosts assign more large parties to a high sales ability waiter to maximize the total sales performance because our sales ability is measured in terms of per person average sales (PPA). Interpreting the coefficients, the effect size is relatively small compared to the mean absolute value of *HrTableDiff*. For average party size that is one standard deviation below the sample mean, the coefficient of *SalesSkill* is approximately $-0.0774+0.037\times(2.04-0.75)\approx -0.02$. That means a waiter having one standard deviation of higher sales ability may receive $0.91\times 0.02 \approx 0.002$ more tables per hour, which is only about $0.002/0.8 \approx 0.002 = 0.2\%$ of mean |*HrTableDiff*|. This effect will be even smaller for average party size that is at the sample mean, since the coefficient of *SalesSkill* becomes approximately $-0.0774+0.037\times 2.04 \approx -0.002$. Hence, waiters' sales skill does not seem to substantially affect hosts' table assignment decisions.

To summarize, despite the heterogeneous effects of store demand and average party size, the effect sizes are relatively small. In addition, within the ranges of the two moderators, respecitvely, the waiter-related factors mostly have the same effect direction as the results without the moderators. For these reasons, we elect to interpret our results in Table 3 as our primary findings, and argue that hosts tend to assign more tables to waiters who have low contemporaneous workload and high speed skills.

Table 5: Moderating Effects of Store Demand and Party Size

	Column 1: Model 5		Column 2: Model 6
TableLoad	-0.1015**	*TableLoad*	0.3433***
	(0.0323)		(0.0319)
TotalTables	0.0045	*AvgPartySize*	0.3045***
	(0.0025)		(0.0295)
TableLoad×*TotalTables*	0.0003	*TableLoad*×*AvgPartySize*	-0.2052***
	(0.0010)		(0.0153)
SpeedSkill	0.0030**	*SpeedSkill*	-0.0149***
	(0.0011)		(0.0010)
SpeedSkill×*TotalTables*	-0.0005***	*SpeedSkill*×*AvgPartySize*	0.0047***
	(0.0000)		(0.0004)
SalesSkill	-0.0020	*SalesSkill*	-0.0774***
	(0.0043)		(0.0050)
SalesSkill×*TotalTables*	0.0003	*SalesSkill*×*AvgPartySize*	0.0370***
	(0.0003)		(0.0022)
WaiterFE	Yes	*WaiterFE*	Yes
HostFE	Yes	*HostFE*	Yes
Controls	Yes	*Controls*	Yes
Observations	595,156	Observations	595,156
Prob>Chi-sq	<.001	Prob>Chi-sq	<.001

1. Standard errors are shown in parentheses. 2. *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$.

7 Performance Implications

7.1 Inverted-U-Shaped Effects of Dispersion of Table Assignments on Performance

In the previous sections, we found evidence that hosts deviate from the rotation rule that they are explicitly instructed to follow, assigning unequal quantities of tables to each waiter based on their workload and speed skills. How this deviation from the RR rule and the resultant dispersion of table assignments among waiters affect the total sales performance in a restaurant are important empirical questions for both academics and practitioners who face increasing pressure in a competitive industry with a low profit margin of 3% to 9% (Mill, 2006). We hypothesize an inverted-U-shaped relationship between the dispersion of table assignments and the total sales in Section 3. To examine this hypothesis, we employ the following model:

$$\log(HrTotalSales_{kt}) = \beta_0 + \beta_1 HrTableGini_{kt} + \beta_2 HrTableGini_{kt}^2 + \beta_3 Controls_{kt} + \varepsilon_{kt}. \tag{7}$$

In this model, $HrTableGini_{kt}$ is the mean-centered Gini Coefficient of the number of tables assigned to each waiter in restaurant k during hour t (the original Gini Coefficient has a mean of 0.164 with a standard deviation of 0.095). The Gini Coefficient is widely used in social sciences as an inequality measure in a distribution (e.g., Yitzhaki, 1979; Lambert and Aronson, 1993). It is a scale-free measure, independent of the total number of tables assigned in an hour. A Gini Coefficient of zero indicates an equal distribution of tables during an hour (i.e., strict RR rule), while a value of one suggests the maximal inequality with all tables allocated to one waiter. We also considered the coefficient of variation of the number of tables assigned as an alternative dispersion measure, which yielded congruent results[3]. In this equation, if β_2 turns out to be significant and negative and $-\beta_1/2\beta_2$ (the critical point of the quadratic function) lies in the range of the *HrTableGini,* we will have evidence for an inverted-U-shaped relationship between *HrTableGini* and the dependent variable *HrTotalSales* (the total sales during hour t in restaurant k, mean = \$375, std = \$226). In addition, we conduct various alternative tests of the inverted-U-relationships, including spline regression and estimating the slopes of the high and low sides of *HrTableGini* to find support for our hypothesis.

Furthermore, *Controls* include the same control variables as in Model 3. We log-transform the dependent variable to increase the normality of the errors and consequently change the interpretation of the coefficients into the percentage impact on the dependent variable. Equally important, Model 7 may be subject to an endogneity bias if the performance-seeking managers pressure hosts to adjust their table assignments. In order to address this endogeneity bias, we use the one-week-lagged *HrTableGini* as an instrument for the current *HrTableGini* in the 2SLS estimation because the one-week lagged dispersion measure, similar to our previous instruments, should satisfy both the relevance and exclusion restriction conditions.

[3] We show the results of using coefficient of variation as an alternative measure in the online Appendix.

Finally, *HrTotalSales* is the product between the hourly number of tables seated, $TotalTables_{kt}$ (mean = 12, std =7.2), and the average sales per table $AvgSales_{kt}$ during the same hour. In order to understand the effect of *HrTableGini* on either component of *HrTotalSales*, we use the following two models:

$$\log(TotalTables_{kt}) = \beta_0 + \beta_1 HrTableGini_{kt} + \beta_2 HrTableGini_{kt}^2 + \beta_3 Controls_{kt} + \varepsilon_{kt} \quad (8)$$

$$\log(AvgSales_{kt}) = \gamma_0 + \gamma_1 HrTableGini_{kt} + \gamma_2 HrTableGini_{kt}^2 + \gamma_3 AvgPartySize_{kt} + \gamma_4 Controls_{kt} + \xi_{kt}. \quad (9)$$

In Model 9, we additionally control for average party size per table at restaurant k during hour t (*AvgPartySize*) because party size should be positively associated with sales at that table. Similar to Model 7, we use a quadratic model specification to examine the inverted-U-shaped relationship between *HrTableGini* and the two dependent variables. We also use the one-week-lagged *HrTableGini* as an instrument for the current *HrTableGini* in the 2SLS estimation for the same endogeneity concern as in Model 7.

Table 6 shows the results of the effects of table assignment dispersion on store performance. The coefficient of $HrTableGini^2$ is significant and negative in Model 7 (-64.6557), while the coefficient of *HrTableGini* is significant and positive (4.8382). This result first supports H4 that predicts an inverted-U-shaped relationship between table assignment dispersion and total sales. The inverted-U-shaped hypothesis is also corroborated by the spline regression results shown in column 4 and the results of estimating the slopes of the two sides of *HrTableGini* shown in the online Appendix. We also find congruent inverted-U-shaped results after dropping top and bottom 5%, and 1% of the data, respectively. Second, the critical point of the concavity is approximately equal to $2 \times 4.8382/64.6557 \approx 0.037$ or 0.4 standard deviation above the sample mean. In other words, hosts' deviating from the rotation rule and giving more tables to waiters who have low contemporaneous workload and high speed skills improves the sales when the deviation size is relatively low. However, when the deviation from the RR rule is too high, further deviation from the RR rule may reduce the total sales because of the unbalanced workload, the negative consequences of perceived unfairness, and the disrupted service process flow. We provide empirical evidence about these mechanisms in the online Appendix.

Interpreting the coefficients, increasing *HrTableGini* to the critical point may improve total sales by

$4.8382 \times 0.037 - 64.6557 \times 0.037^2 \approx 9\%$ from the status quo. That is to say, hosts use their local knowledge to adjust the seating rule and improve sales performance in the right direction. When doing the adjustments, however, they still have some unfulfilled potential for sales improvement.In order to understand where this counter-factual 9% sales lift results from, we examine the effect of *HrTableGini* on *Tables* and *AvgSales,* two factors of the total sales. As can be seen, the coefficients of *HrTablesGini*^{2} are consistently significant and negative in both Models 8 and 9 (-59.2106 and -4.1213); the coefficients of *HrTablesGini* are both significant and positive (4.6507 and 0.1174), suggesting *HrTableGini* may have an inverted-U-shaped relationship with either *Tables* or *AvgSales.* Spline regression results shown in Columns 5 and 6 also support the inverted-U-shaped relationships. Moreover, increasing *HrTableGini* by 0.037 (the optimal amount to maximize the total sales) may increase *Tables* by 9% $4.6507 \times 0.037 - 59.2016 \times 0.037^2 \approx 9\%$, and yet reduce *AvgSales* by $0.1174 \times 0.037 - 4.1213 \times 0.037^2 \approx 0.12\%$. Hence, the majority of the sales lift from adjusting the seating rule is due to the increase of the extra tables seated. We expect this result because we find that hosts tend to assign more tables to those servers who have high speed skills instead of high sales skills. Assigning tables to faster workers may efficiently turn the tables, reducing customers' waiting time and avoiding lost sales due to balking - the 9% additional tables implies $9\% \times 12 \approx 1$ extra table on average. This creates opportunities for the restaurants to seat extra parties. For example, when shadowing the hosts, we observed approximately four parties waiting between 7 pm and 8 pm on a weekend. In addition, the host seemed to overestimate the customers' waiting time in order to set a low expectation for customers and to avoid their complaints. A significant number of customers therefore balked. Moreover, we saw hosts sometimes holding the customers in the waiting area in order not to overload the waiters, even when there were empty tables. Had the host used these tables, fewer customers would have balked. To sum up, the majority of the sales lift from adjusting the seating rule currently results from turning tables. However, there is untapped potential to improve sales per table when seating customers. Overall there is growth potential for the restaurants by increasing table turns and average sales per table.

Table 6: Implications of Dispersion of Table Assignments

	Column 1:	Column 2:	Column 3:		Column 4:	Column 5:	Column 6:
	Model 7	Model 8	Model 9		Model 7	Model 8	Model 9
					with Spline	with Spline	with Spline
					Regression	Regression	Regression
HrTableGini	4.8382***	4.6507***	0.1174*	*HrTableGini_1*	16.9648***	15.7561***	0.9566***
	(0.2630)	(0.2399)	(0.0507)		(0.6603)	(0.6030)	(0.1104)
HrTableGini^2	-64.6557***	-59.2106***	-4.1213***	*HrTableGini_2*	-10.6190***	-9.5047***	-0.9182***
	(3.3335)	(3.0408)	(0.7016)		(0.9827)	(0.8974)	(0.1647)
AvgPartySize			0.3603***	*AvgPartySize*			0.3603***
			(0.0012)				(0.0012)
Controls	Yes	Yes	Yes	*Controls*	Yes	Yes	Yes
Observations	131,682	131,682	131,360	Observations	131,682	131,682	131,360
Prob>Chi-sq	<.001	<.001	<.001	Prob>Chi-sq	<.001	<.001	<.001

1. Standard errors are shown in parentheses. 2. *$p\leq$ .05, **$p\leq$ .01, ***$p\leq$ .001.

7.2 Counter-factual Analysis of Assigning More Tables to High Sales Skill Waiters

Our results suggest that hosts' adjusting the rotation seating rule increases sales performance in the right direction. However, there are still opportunities for further improvement because the optimal deviation *HrTablesGini* is still above the sample mean. On the one hand, hosts may strengthen the magnitude of their current adjustment decisions, and assign even more tables to those waiters that have low workload and high speed skills. On the other hand, they may further start to unlock the potential of those waiters who have higher sales skills and assign them more tables. Below we run four counter-factual analyses to estimate the potential sales improvement from prioritizing high sales skill waiters.

We first assume that capacity is equal to four, five or six tables per waiter per hour respectively, because six tables is the maximum in our data. Then, every hour we fill the waiter having the highest sales skills to their capacity, followed by the waiter having the second highest sales skills, and repeat this assignment rule until we finish assigning all the tables. If the total number of tables per hour occasionally exceed the total capacity (less than 9% of the time when average capacity is four tables, less than 2.5% of time when average capacity is five tables, and less than 0.7% when average capacity is six tables), we assign the extra tables to the highest sales skill waiter. After that, we compute the weighted average of the number of customers by the number of tables assigned, which represents the counter-factual number of customers per waiter. From these counter-factual numbers of customers, we then subtract the actual number of customers served. Finally, we multiply these differences by the waiter's sales skills (a fixed effect of per person average) estimated in Model 2 to compute the counter-factual sales differences.

Table 7 presents the results of our counter-factual analyses of prioritizing sales skills. The average sales difference ranges from $1.56 per waiter per hour (1.8% sales lift) when we ensure a capacity of four tables and a minimum of one table to $2.78 (3.2% sales lift) when we assume a capacity of six tables with no minimum table requirement. These estimates are a significant sales lift for a low profit margin industry, without changing other labor decisions, such as staffing or scheduling.

In practice, prioritizing some workers may create a perception of unfairness among workers and cause frustration. Nevertheless, carefully prioritizing high skill waiters may motivate low skill ones to improve their skills in order to receive more tables and earn more income (Lazear and Rosen, 1981; Lazear, 1989). In other words, a well-executed dispersion of routing jobs among workers may produce a Pareto improvement. For example, Netessine and Yakubovich (2012) show that some restaurants track waitstaff performance and reward high performance waiters with more tables and preferred schedules to incentivize the entire waitstaff. Finally, our counter-factual analyses in both Tables 6 and 7 show the value of routing decisions, and therefore have implications for hiring, training and compensating those seemingly unimportant workers such as hosts, who actually can work around a suboptimal instructed rule and significantly influence organizational performance. This is consistent with Bowman (1963)'s classic Managerial Coefficient Theory that states front line workers may possess relevant information that is not accounted for in the implemented rule.

Table 7: Counter-factual Analyses of Prioritizing Sales Skills

	Analysis I	Analysis II	Analysis III	Analysis IV
Capacity	4	5	5	6
Minimum Tables	1	1	0	0
Average Sales Difference (per waiter per hour)	$1.56 (1.8%)	$2.06 (2.37%)	$2.11 (2.43%)	$2.78 (3.2%)

8 Discussion and Conclusion

In this paper, we find that hosts adjust their given RR rule, which is supposed to ensure fairness to servers and smooth the service flow. Instead, they direct more incoming parties to those waiters that have low contemporaneous workload or high speed skills than the RR rule suggests, but they do not seat more incoming parties to those waiters that have high sales skills. The inequality of the table assignment is moderated by store demand and waiter's average party size. As store demand increases, hosts tend to give more tables to faster waiters. In addition, hosts are inclined to assign larger parties to waiters having low contemporaneous workload, low speed ability and high sales ability. Furthermore, we find that as the inequality of the table assignments (measured in terms of the Gini Coefficient of the number of tables assigned to the waiters during the same hour) increases, total store sales first increase and then decrease. In other words, there is an inverted-U-shaped relationship between the deviation from the RR rule and the total sales. In particular, the optimal Gini Coefficient is approximately 0.4 standard deviation above the sample mean. Prioritizing fast waiters to achieve this optimal deviation may increase total sales by approximately 9%, from increasing throughput, not changing average sales per table. We anticipate that our results should be qualitatively generalizable to other hosts and restaurants. However, workers' improvisation behavior and its impact depend on their competence (Secchi et al., 2019). More experienced or better trained hosts at higher-end restaurants should be more likely to improvise and have even stronger sales impact than the hosts in our focal causal restaurant chain.

Our paper makes three primary contributions to the service operations literature. First, to the best of our knowledge, our paper is the first to empirically examine how individuals make routing decisions. It is particularly related to a growing stream of routing research that explicitly models the speed/quality trade-off and the efficiency/fairness trade-off (e.g., De Véricourt and Zhou, 2005; Mandelbaum et al., 2012; Mehrotra et al., 2012; Ward and Armony, 2013; Zhan and Ward, 2013). Our empirical findings support the underlying structure or principle of this stream of research by showing that hosts follow a priority rule that allocates more work to some types of waiters, even though this rule is different from what they are instructed to follow. Second, our paper is the first to use observational data to study how individuals adjust a given routing rule in an applied setting. It contributes to the literature that researches how humans actually make operational decisions, which may deviate from predefined policies. Much of this literature initially took place in the laboratory as an excellent way to control for numerous factors. However, as the research continues to develop, there is an opportunity to move into the field to gain external validity and to understand in more depth the decisions that individuals make (Cui et al., 2015; Elmaghraby et al., 2015; Van Donselaar et al., 2010). This, in turn, creates new opportunities for lab-based work to gain greater insight. Individual routing decisions are ripe for such exploration, although previous studies have not investigated it. Finally, our paper shows the value of frontline workers (i.e., the hosts), who apply relevant information that is not accounted for in their given routing rule. It answers to the call by Boudreau et al. (2003) to explore the role of people in operations. People-centric operations is an important area to continue to investigate given the limits of what we know now and the changes taking place (e.g., artificial intelligence, Economist, 2016), that will alter how humans affect operations.

Our research also has important and direct implications for managers. We suggest that that managers should be first aware of hosts' discretionary routing preferences, and applaud them for adjusting the RR rule in the right direction. In addition, they should consider encouraging the hosts to intensify their prioritization and start to prioritize those waiters having high sales skills, all else being equal. Nevertheless, we caution that prioritization or deviating from the RR rule, is a double-edged sword. On the one hand, appropriate prioritization may improve system performance; on the other hand, excessive uneven routing to some servers can be detrimental. Future work should examine not only the short-term effects of these choices, as we have done here, but also the longer term effects. For example, what happens over time in a restaurant that routes work unevenly. Does it lead to an overall improvement, or instead, does it lead to higher turnover and inadvertently lower performance? Finally, our research also reminds managers to pay more attention to the host position and other similar positions, which are often neglected but have great importance for company performance.

We conclude by highlighting important limitations of our findings and outlining opportunities for future research. First, ideally we would have waiting time information. Unfortunately, almost no advanced technology is available to systematically track the whole restaurant process. Some startups are just emerging to provide such technology, which may create opportunities for future research to analyze the impact of seating rules on customer waiting time and potential lost sales. Second, our data do not have information about the number of table "tops" (i.e., table size). It is possible that hosts may occasionally adjust the seating sequence to match the party size with the table size. We control for table fixed effects as a robustness check to support our main results, alleviating such omitted variable bias concern. Nevertheless, it would be interesting to study how hosts take this table size factor into their seating decisions. Third, our data do not capture customer history. Customer familiarity with a host or with a waiter may affect table assignment. To alleviate this concern, we adjust for waiter's and hosts' fixed effects. In addition, such a familiarity effect may be limited in our empirical setting because, unlike fine-dining, casual dining experiences fast staff turnover. Fourth, we do not observe the actual seating time or the system prompts about the actual round robin rule. We use hourly aggregate level data to address this issue. More research can analyze actual seating time and how hosts may deviate from what the system suggests (e.g., Van Donselaar et al., 2010). Finally, we do not study how waiters may change their performance in response to hosts' seating decisions. However, there are many fruitful opportunities to study how workers endogenize their efforts when their work assignment is different from their peers (see Geng et al., 2014; Valentine, 2018 for some initial work in this direction).

Although in this paper, we do not study all the factors that may affect hosts' discretionary routing decision process, we sincerely hope that our study will serve as a useful first step because our literature lacks empirical knowledge about how individuals make routing decisions.

References

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

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

Angrist, J., A. B. Krueger. 1994. Why do World War II veterans earn more than nonveterans? *Journal of Labor Economics* **12**(1) 74–97.

Argote, L., D. Epple. 1990. Learning curves in manufacturing. *Science* **247**(4945) 920–924.

Armony, M., A. R. Ward. 2010. Fair dynamic routing in large-scale heterogeneous-server systems. *Operations Research* **58**(3) 624–637.

Armony, Mor. 2005. Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems* **51**(3-4) 287–329.

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.

Bertsimas, D., V. F. Farias, N. Trichakis. 2012. On the efficiency-fairness trade-off. *Management Science* **58**(12) 2234–2250.

Bloom, N., J. Van Reenen. 2007. Measuring and explaining management practices across firms and countries. *Quarterly Journal of Economics* **122**(4) 1351–1408.

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.

Bowman, E. H. 1963. Consistency and optimality in managerial decision making. *Management Science* **9**(2) 310–321.

Chan, T. Y., J. Li, L. Pierce. 2014. Compensation and peer effects in competing sales teams. *Management Science* **60**(8) 1965–1984.

Cohen-Charash, Y., P. E. Spector. 2001. The role of justice in organizations: A meta-analysis. *Organizational behavior and human decision processes* **86**(2) 278–321.

Colquitt, J. A., D. E. Conlon, M. J. Wesson, C. Porter, K. Y. Ng. 2001. Justice at the millennium: a meta-analytic review of 25 years of organizational justice research. *Journal of applied psychology* **86**(3) 425.

Croson, R., K. Donohue. 2006. Behavioral causes of the bullwhip effect and the observed value of inventory information. *Management Science* **52**(3) 323–336.

Cui, R., G. Allon, A. Bassamboo, J. A. Van Mieghem. 2015. Information sharing in supply chains: An empirical and theoretical valuation. *Management Science* **61**(11) 2803–2824.

De Véricourt, F., Y. P. Zhou. 2005. Managing response time in a call-routing problem with service failure. *Operations Research* **53**(6) 968–981.

Deo, S., I. Gurvich. 2011. Centralized vs. decentralized ambulance diversion: A network perspective. *Management Science* **57**(7) 1300–1319.

Economist, The. 2016. Special report on artificial intelligence: Automation and anxiety. *The Economist* .

Elmaghraby, W., W. Jank, S. Zhang, I. Z. Karaesmen. 2015. Sales force behavior, pricing information, and pricing decisions. *Manufacturing & Service Operations Management* .

Frostig, E., B. Levikson. 1999. Optimal routing of customers to two parallel heterogeneous servers: The case of ihr service times. *Operations research* **47**(3) 438–444.

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

Geng, X., W. T. Huh, M. Nagarajan. 2014. Fairness among servers when capacity decisions are endogenous. *Production and Operations Management* .

Gopalan, R., K. T. Talluri. 1998. The aircraft maintenance routing problem. *Operations Research* **46**(2) 260–271.

Gurvich, I., W. Whitt. 2009. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing & Service Operations Management* **11**(2) 237–253.

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.

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

Kostami, V., S. Rajagopalan. 2013. Speed-quality trade-offs in a dynamic model. *Manufacturing & Service Operations Management* **16**(1) 104–118.

Kremer, M., B. Moritz, E. Siemsen. 2011. Demand forecasting behavior: System neglect and change detection. *Management Science* **57**(10) 1827–1843.

Lambert, P.J., J.R. Aronson. 1993. Inequality decomposition analysis and the gini coefficient revisited. *The Economic Journal* 1221–1227.

Lazear, E. P. 1989. Pay equality and industrial politics. *Journal of Political Economy* **97**(3) 561–580.

Lazear, E. P., S. Rosen. 1981. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* **89**(5) 841–864.

Lee, Hau Leung, Morris A Cohen. 1985. Multi-agent customer allocation in a stochastic service system. *Management Science* **31**(6) 752–763.

Mandelbaum, A., P. Momcilovic, Y. Tseytlin. 2012. On fair routing from emergency departments to hospital wards: Qed queues with heterogeneous servers. *Management Science* **58**(7) 1273–1291.

Mandelbaum, A., A. L. Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research* **52**(6) 836–855.

Mas, A., E. Moretti. 2009. Peers at work. *American Economic Review* **99**(1) 112–145.

Mehrotra, V., K. Ross, G. Ryder, Y. P. Zhou. 2012. Routing to manage resolution and waiting time in call centers with heterogeneous servers. *Manufacturing & service operations management* **14**(1) 66–81.

Mill, R. C. 2006. *Restaurant management: Customers, operations, and employees*. Upper Saddle River, NJ: Prentice Hall.

Narasimhan, R., S. Narayanan, R. Srinivasan. 2013. An investigation of justice in supply chain relationships and their performance impact. *Journal of Operations Management* **31**(5) 236–247.

Narayanan, S., S. Balasubramanian, J. M. Swaminathan. 2009. A matter of balance: Specialization, task variety, and individual learning in a software maintenance environment. *Management Science* **55**(11) 1861–1876.

Netessine, S., V. Yakubovich. 2012. The darwinian workplace. *Harvard business review* **90**(5) 25–6.

Ovchinnikov, A., B. Moritz, B. F. Quiroga. 2015. How to compete against a behavioral newsvendor. *Production and Operations Management* .

Parekh, A.K., R.G. Gallager. 1993. A generalized processor sharing approach to flow control in integrated services networks: the single-node case. *IEEE/ACM Transactions on Networking (TON)* **1**(3) 344–357.

Schweitzer, M. E., G. P. Cachon. 2000. Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science* **46**(3) 404–420.

Secchi, E., A. Roth, R Verma. 2019. The impact of service improvisation competence on customer satisfaction: evidence from the hospitality industry. *Production and Operations Management* **28**(6) 1329–1346.

Shafer, S. M., D. A. Nembhard, M. V. Uzumeri. 2001. The effects of worker learning, forgetting, and heterogeneity on assembly line productivity. *Management Science* 1639–1653.

Shah, R., G. P. Ball, S. Netessine. 2016. Plant operations and product recalls in the automotive industry: An empirical investigation. *Management Science* **63**(8) 2439–2459.

Siebert, W. S., N. Zubanov. 2010. Management economics in a large retail company. *Management Science* **56**(8) 1398–1414.

Song, H., A. L. Tucker, K. L. Murrell. 2015. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science* **61**(12) 3032–3053.

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

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

Sterman, J. D. 1989. Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management science* **35**(3) 321–339.

Tan, T. F., S. Netessine. 2014. When does the devil make work? an empirical study of the impact of workload on worker productivity. *Management Science* .

Tan, T. F., S. Netessine. 2015. When you work with a super man, will you also fly? an empirical study of the impact of coworkers on performance. Working Paper.

Tezcan, T., J. Zhang. 2014. Routing and staffing in customer service chat systems with impatient customers. *Operations research* **62**(4) 943–956.

Thaler, R. 1981. Some empirical evidence on dynamic inconsistency. *Economics letters* **8**(3) 201–207.

Tucker, A. 2015. The impact of workaround difficulty on frontline employees response to operational failures: A laboratory experiment on medication administration. *Management Science* **62**(4) 1124–1144.

Tucker, A. L. 2007. An empirical study of system improvement by frontline employees in hospital units. *Manufacturing & Service Operations Management* **9**(4) 492–505.

Valentine, M. 2018. When equity seems unfair: The role of justice enforceability in temporary team coordination. *Academy of Management Journal* (Forthcoming) amj–2016.

Van Donselaar, K. H., V. Gaur, T. Van Woensel, R. A. C. M. Broekmeulen, J. C. Fransoo. 2010. Ordering behavior in retail stores and implications for automated replenishment. *Management Science* **56**(5) 766–784.

Walker, J. R. 2007. *The Restaurant, Study Guide: From Concept to Operation*. New York City, NY: John Wiley & Sons.

Ward, A. R., M. Armony. 2013. Blind fair routing in large-scale service systems with heterogeneous customers and servers. *Operations Research* **61**(1) 228–243.

Wooldridge, J. M. 2002. *Econometric analysis of cross section and panel data*. The MIT press.

Yitzhaki, S. 1979. Relative deprivation and the gini coefficient. *The Quarterly Journal of Economics* 321–324.

Zhan, D., A. R. Ward. 2013. Threshold routing to trade off waiting and call resolution in call centers. *Manufacturing & Service Operations Management* **16**(2) 220–237.