MANAGEMENT SCIENCE
Vol. 59, No. 8, August 2013, pp. 1743–1763
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.1120.1686

Measuring the Effect of Queues on Customer Purchases

Yina Lu
Decision, Risk, and Operations Division, Columbia Business School, Columbia University, New York, New York 10027, yl2494@columbia.edu

Andrés Musalem
Fuqua School of Business, Duke University, Durham, North Carolina 27708, andres.musalem@duke.edu

Marcelo Olivares
Decision, Risk, and Operations Division, Columbia Business School, Columbia University, New York, New York 10027; and Department of Industrial Engineering, University of Chile, Santiago, Chile 8370439, molivares@columbia.edu

Ariel Schilkrut
SCOPIX, Burlingame, California 94010, ariel.schilkrut@scopixsolutions.com

We conduct an empirical study to analyze how waiting in queue in the context of a retail store affects customers' purchasing behavior. Our methodology combines a novel data set with periodic information about the queuing system (collected via video recognition technology) with point-of-sales data. We find that waiting in queue has a nonlinear impact on purchase incidence and that customers appear to focus mostly on the length of the queue, without adjusting enough for the speed at which the line moves. An implication of this finding is that pooling multiple queues into a single queue may increase the length of the queue observed by customers and thereby lead to lower revenues. We also find that customers' sensitivity to waiting is heterogeneous and negatively correlated with price sensitivity, which has important implications for pricing in a multiproduct category subject to congestion effects.

Key words: queuing; service operations; retail; choice modeling; empirical operations management; operations/marketing interface; Bayesian estimation; service quality
History: Received May 24, 2011; accepted August 10, 2012, by Martin Lariviere, operations management. Published online in *Articles in Advance* April 22, 2013.

1. Introduction

Capacity management is an important aspect in the design of service operations. These decisions involve a trade-off between the costs of sustaining a service level standard and the value that customers attach to it. Most work in the operations management literature has focused on the first issue developing models that are useful to quantify the costs of attaining a given level of service. Because these operating costs are more salient, it is frequent in practice to observe service operations rules designed to attain a quantifiable target service level. For example, a common rule in retail stores is to open additional checkouts when the length of the queue surpasses a given threshold. However, there isn't much research focusing on how to choose an appropriate target service level. This requires measuring the value that customers assign to objective service level measures and how this translates into revenue. The focus of this paper is to measure the effect of service levels—in particular, customers waiting in queue—on actual customer purchases, which can be used to attach an economic value to customer service.

Lack of objective data is an important limitation to study empirically the effect of waiting on customer behavior. A notable exception is call centers, where some recent studies have focused on measuring customer impatience while waiting on the phone line (Gans et al. 2003). Instead, our focus is to study *physical* queues in services, where customers are physically present at the service facility during the wait. This type of queue is common, for example, in retail stores, banks, amusement parks, and healthcare delivery. Because objective data on customer service are typically not available in these service facilities, most previous research relies on surveys to study how customers' *perceptions* of waiting affect their *intended* behavior. However, previous work has also shown that customer perceptions of service do not necessarily match with the actual service level received, and purchase intentions do not always translate into actual revenue (e.g., Chandon et al. 2005). In contrast, our work uses objective measures of actual service collected through a novel technology—digital imaging with image recognition—that tracks operational metrics such as the number of customers waiting in

line. We develop an econometric framework that uses these data together with point-of-sales (POS) information to estimate the impact of customer service levels on purchase incidence and choice decisions. We apply our methodology using field data collected in a pilot study conducted at the deli section of a big-box supermarket. An important advantage of our approach over survey data is that the regular and frequent collection of the store operational data allows us to construct a large panel data set that is essential to identifying each customer's sensitivity to waiting.

There are two important challenges in our estimation. A first issue is that congestion is highly dependent on store traffic, and therefore periods of high sales are typically concurrent with long waiting lines. Consequently, we face a reverse causality problem: whereas we are interested in measuring the causal effect of waiting on sales, there is also a reverse effect whereby spikes in sales generate congestion and longer waits. The correlation between waiting times and aggregate sales is a combination of these two competing effects and therefore cannot be used directly to estimate the causal effect of waiting on sales. The detailed panel data with purchase histories of individual customers is used to address this issue.

Using customer transaction data produces a second estimation challenge. The imaging technology captures snapshots that describe the queue length and staffing level at specific time epochs but does not provide an exact measure of what is observed by each customer (technological limitations and consumer privacy issues preclude us from tracking the identity of customers in the queue). A rigorous approach is developed to infer these missing data from periodic snapshot information by analyzing the transient behavior of the underlying stochastic process of the queue. We believe this is a valuable contribution that will facilitate the use of periodic operational data in other studies involving customer transactions obtained from POS information.

Our model also provides several metrics that are useful for the management of service facilities. First, it provides estimates on how service levels affect the effective arrivals to a queuing system when customers may balk. This is a necessary input to set service and staffing levels optimally balancing operating costs against lost revenue. In this regard, our work contributes to the stream of empirical research related to retail staffing decisions (e.g., Fisher et al. 2009, Perdikaki et al. 2012). Second, it can be used to identify the relevant visible factors in a physical queuing system that drive customer behavior, which can be useful for the design of a service facility. Third, our models provide estimates of how the performance of a queuing system may affect how customers substitute among alternative products or services accounting for heterogeneous customer preferences. Finally, our methodology can be used to attach a dollar value to the cost of waiting experienced by customers and to segment customers based on their sensitivity to waiting.

In terms of our results, our empirical analysis suggests that the number of customers in the queue has a significant impact on the purchase incidence of products sold in the deli, and this effect appears to be nonlinear and economically significant. Moderate increases in the number of customers in queue can generate sales reduction equivalent to a 5% price increase. Interestingly, the service capacity—which determines the speed at which the line moves—seems to have a much smaller impact relative to the number of customers in line. This is consistent with customers using the number of people waiting in line as the primary visible cue to assess the expected waiting time. This empirical finding has important implications for the design of the service facility. For example, we show that pooling multiple queues into a single queue with multiple servers may lead to more customers walking away without purchasing and therefore lower revenues (relative to a system with multiple queues). We also find significant heterogeneity in customer sensitivity to waiting, and that the degree of waiting sensitivity is negatively correlated with customers' sensitivity to price. We show that this result has important implications for pricing decisions in the presence of congestion and, consequently, should be an important element to consider in the formulation of analytical models of waiting systems.

2. Related Work

In this section, we provide a brief review of the literature studying the effect of waiting on customer behavior and its implications for the management of queues. Extensive empirical research using experimental and observational data has been done in the fields of operations management, marketing, and economics. We focus this review on a selection of the literature that helps us to identify relevant behavioral patterns that are useful in developing our econometric model (described in §3). At the same time, we also reference survey articles that provide a more exhaustive review of different literature streams.

Recent studies in the service engineering literature have analyzed customer transaction data in the context of call centers. See Gans et al. (2003) for a survey on this stream of work. Customers arriving to a call center are modeled as a Poisson process where each arriving customer has a "patience threshold": one abandons the queue after waiting more than his patience threshold. This is typically referred to as the Erlang-A model or the $M/M/c+G$,

where G denotes the generic distribution of the customer patience threshold. Brown et al. (2005) estimate the distribution of the patience threshold based on call-center transactional data and use it to measure the effect of waiting time on the number of lost (abandoned) customers.

Customers arriving to a call center typically do not directly observe the number of customers ahead in the line, so the estimated waiting time may be based on delay estimates announced by the service provider or their prior experience with the service (Ibrahim and Whitt 2011). In contrast, for physical customer queues at a retail store, the length of the line is observed and may become a visible cue affecting their perceived waiting time. Hence, queue length becomes an important factor in customers' decision to join the queue, which is not captured in the Erlang-A model. In these settings, arrivals to the system can be modeled as a Poisson process where a fraction of the arriving customers may *balk*—that is, not join the queue—depending on the number of people already in queue (see Gross et al. 2008, Chap. 2.10). Our work focuses on estimating how visible aspects of physical queues, such as queue length and capacity, affect choices of arriving customers, which provides an important input to normative models.

Png and Reitman (1994) empirically study the effect of waiting time on the demand for gas stations and identify service time as an important differentiating factor in this retail industry. Their estimation is based on aggregate data on gas station sales and uses measures of a station's capacity as a proxy for waiting time. Allon et al. (2011) study how service time affects demand across outlets in the fast food industry, using a structural estimation approach that captures price competition across outlets. Both studies use aggregate data from a cross-section of outlets in local markets. The data for our study are more detailed because they use individual customer panel information and periodic measurements of the queue, but it is limited to a single service facility. None of the aforementioned papers examine heterogeneity in waiting sensitivity at the individual level as we do in our work.

Several empirical studies suggest that customer responses to waiting time are not necessarily linear. Larson (1987) provides anecdotal evidence of nonlinear customer disutility under different service scenarios. Laboratory and field experiments have shown that customer's perceptions of waiting are important drivers of dissatisfaction and that these perceptions may be different from the actual (objective) waiting time, sometimes in a nonlinear pattern (e.g., Davis and Vollmann 1993, Berry et al. 2002, Antonides et al. 2002). Mandelbaum and Zeltyn (2004) use analytical queuing models with customer impatience to explain nonlinear relationships between waiting time and customer abandonment. Indeed, in the context of call-center outsourcing, the common use of service level agreements based on delay thresholds at the upper tail of the distribution (e.g., 95% of the customers wait less than two minutes) is consistent with nonlinear effects of waiting on customer behavior (Hasija et al. 2008).

Larson (1987) provides several examples of factors that affect customers' perceptions of waiting, such as (1) whether the waiting is perceived as socially fair, (2) whether the wait occurs before or after the actual service begins, and (3) feedback provided to the customer on waiting estimates and the root causes generating the wait, among other examples. Berry et al. (2002) provide a survey of empirical work testing some of these effects. Part of this research has used controlled laboratory experiments to analyze factors that affect customers perceptions of waiting. For example, the experiments by Hui and Tse (1996) suggest that queue length has no significant impact on service evaluation in short-wait conditions, although it has a significant impact on service evaluation in long-wait conditions. Janakiraman et al. (2011) use experiments to analyze customer abandonments and propose two competing effects that explain why abandonments tend to peak at the midpoint of waits. Hui et al. (1997) and Katz et al. (1991) explore several factors, including music and other distractions, that may affect customers' perception of waiting time.

In contrast, our study relies on field data to analyze the effect of queues on customer purchases. Much of the existing field research relies on surveys to measure objective and subjective waiting times, linking these to customer satisfaction and intentions of behavior. For example, Taylor (1994) studies a survey of delayed airline passengers and finds that delay decreases service evaluations by invoking uncertainty and anger affective reactions. Deacon and Sonstelie (1985) evaluate customers' time value of waiting based on a survey on gasoline purchases. Although surveys are useful to uncover the behavioral process by which waiting affects customer behavior and the factors that mediate this effect, they also suffer from some disadvantages. In particular, there is a potential sample selection because nonrespondents tend to have a higher opportunity cost for their time. In addition, several papers report that customer purchase intentions do not always match actual purchasing behavior (e.g., Chandon et al. 2005). Moreover, relying on surveys to construct a customer panel data set with the required operational data is difficult (all the referenced articles use a cross-section of customers). Our work uses measures of not only actual customer purchases, but also operational drivers of waiting time (e.g., queue length and capacity at the time of each customer visit) to construct a panel with objective metrics of purchasing behavior and waiting. Our

approach, however, is somewhat limited for studying some of the underlying behavioral process driving the effect of waiting time.

Several other studies use primary and secondary observational data to measure the effect of service time on customer behavior. Forbes (2008) analyzes the impact of airline delays on customer complaints, showing that customer expectations play an important role mediating this effect. Campbell and Frei (2011) study multiple branches of a bank, providing empirical evidence that teller waiting times affect customer satisfaction and retention. Their empirical study reveals significant heterogeneity in customer sensitivity to waiting time, some of which can be explained through demographics and the intensity of competition faced by the branch. Akşin et al. (2013) model callers' abandonment decisions as an optimal stopping problem in a call-center context and find heterogeneity in callers' waiting behavior. Our study also looks at customer heterogeneity in waiting sensitivity, but in addition we relate this sensitivity to customers' price sensitivity. This association between price and waiting sensitivity has important managerial implications; for example, Afèche and Mendelson (2004) and Afanasyev and Mendelson (2010) show that it plays an important role for setting priorities in queue and it affects the level of competition among service providers. Section 5 discusses other managerial implications of this price/waiting sensitivity relationship in the context of category pricing.

Our study uses discrete choice models based on random utility maximization to measure substitution effects driven by waiting. The same approach was used by Allon et al. (2011), who incorporated waiting time factors into customers' utility using a multinomial logit (MNL) model. We instead use a random coefficient MNL, which incorporates heterogeneity and allows for more flexible substitution patterns (Train 2003). The random coefficient MNL model has also been used in the transportation literature to incorporate the value of time in consumer choice (e.g., Hess et al. 2005).

Finally, all of the studies mentioned so far focus on settings where waiting time and congestion generate disutility to customers. However, there is theory suggesting that longer queues could create value to a customer. For example, if a customers' utility for a good depends on the number of customers that consume it (as with positive network externalities), then longer queues could attract more customers. Another example is given by herding effects, which may arise when customers have asymmetric information about the quality of a product. In such a setting, longer queues provide a signal of higher value to uninformed customers, making them more likely to join the queue (see Debo and Veeraraghavan 2009 for several examples).

3. Estimation

This section describes the data and models used in our estimation. The literature review of §2 provides several possible behavioral patterns that are included in our econometric specification: (1) the effect of waiting time on customer purchasing behavior may be nonlinear, such that customers' sensitivity to a marginal increase in waiting time may vary at different levels of waiting time; (2) the effect may not be monotone (for example, although more anticipated waiting is likely to negatively affect customers' purchase intentions, herding effects could potentially make longer queues attractive to customers); (3) customer purchasing behavior is affected by perceptions of waiting time, which may be formed based on the observed queue length and the corresponding staffing level; (4) customers' sensitivity to waiting time may be heterogeneous and possibly related to demographic factors, such as income or price sensitivity.

Subsection 3.1 describes the data used in our empirical study, which motivates the econometric framework developed in the rest of the section. Subsection 3.2 describes an econometric model to measure the effect of queues on purchase incidence. It uses a flexible functional form to measure the effect of the queue on purchasing behavior that permits potential nonlinear and nonmonotone effects. Different specifications are estimated to test for factors that may affect customers' perceptions of waiting. Subsection 3.3 describes how to incorporate the periodic queue information contained in the snapshot data into the estimation of this model. Subsection 3.4 conducts a simulation study to validate this estimation methodology. Subsection 3.5 develops a discrete choice model that captures additional factors not incorporated into the purchase incidence model, including substitution among products, prices, promotions, and state-dependent variables that affect purchases (e.g., household inventory). This choice model is also used to measure heterogeneity in customer sensitivity to waiting.

3.1. Data

We conducted a pilot study at the deli section of a supercenter located in a major metropolitan area in Latin America. The store belongs to a leading supermarket chain in this country and is located in a working-class neighborhood. The deli section sells about eight product categories, most of which are fresh cold-cuts sold by the pound.

During a pilot study running from October 2008 to May 2009 (approximately seven months), we used digital snapshots analyzed by image recognition technology to periodically track the number of people waiting at the deli and the number of sales associates

Figure 1 Example of a Deli Snapshot Showing the Number of Customers Waiting (Left) and the Number of Employees Attending (Right)

Source. Courtesy of SCOPIX.

serving it. Snapshots were taken periodically every 30 minutes during the open hours of the deli, from 9 A.M. to 9 P.M. on a daily basis. Figure 1 shows a sample snapshot that counts the number of customers waiting (left panel) and the number of employees attending customers behind the deli counter (right panel). Throughout this paper, we denote the length of the deli queue at snapshot t by Q_t and the number of employees serving the deli by E_t.

During peak hours, the deli uses numbered tickets to implement a first-come, first-served priority in the queue. The counter displays a visible panel intended to show the ticket number of the last customer attended by a sales associate. This information would be relevant for the purpose of our study to complement the data collected through the snapshots; for example, Campbell and Frei (2011) use ticket-queue data to estimate customer waiting time. However, in our case the ticket information was not stored in the POS database of the retailer, and we learned from other supermarkets that this information is rarely recorded. Nevertheless, the methods proposed in this paper could also be used with periodic data collected via a ticket queue, human inspection, or other data collection procedures.

In addition to the queue and staffing information, we also collected POS data for all transactions involving grocery purchases from January 1, 2008, until the end of the study period. In the market area of our study, grocery purchases typically include bread, and about 78% of the transactions that include deli products also include bread. For this reason, we selected basket transactions that included bread to obtain a sample of grocery-related shopping visits. Each transaction contains checkout data, including a time stamp of the checkout and the stock-keeping units (SKUs) bought along with unit quantities and prices (after promotions). We use the POS data prior to the pilot study period—from January to September of 2008—to calculate metrics employed in the estimation of some our models (we refer to this subset of the data as the *calibration* data).

Using detailed information on the list of products offered at this supermarket, each cold-cut SKU was assigned to a product category (e.g., ham, turkey, bologna, salami, etc.). Some of these cold-cut SKUs include prepackaged products that are not sold by the pound and therefore are located in a different section of the store.[1] For each SKU, we defined an attribute indicating whether it was sold in the deli or prepackaged section. About 29.5% of the transactions in our sample include deli products, suggesting that deli products are quite popular in this supermarket.

An examination on the hourly variation of the number of transactions, queue length, and number of employees reveals the following interesting patterns. In weekdays, peak traffic hours are observed around midday, between 11 A.M. and 2 P.M., and in the evenings, between 6 P.M. and 8 P.M. Although there is some adjustment in the number of employees attending, this adjustment is insufficient, and therefore queue lengths exhibit an hour-of-day pattern similar to the one for traffic. A similar effect is observed for weekends, although the peak hours are different. In other words, congestion generates a positive correlation between aggregate sales and queue lengths, making it difficult to study the causal effect of queues on traffic using aggregate POS data. In our empirical study, detailed *customer transaction* data are used instead to address this problem. More specifically, the supermarket chain in our study operates a popular loyalty program such that more than 60% of the transactions are matched with a loyalty card identification number, allowing us to construct a panel of

[1] This prepackaged section can be seen to the right of customer numbered 1 in the left panel of Figure 1 (top-right corner).

Table 1 Summary Statistics of the Snapshot Data, Point-of-Sales Data, and Loyalty Card Data

	No. of obs.	Mean	Std. dev.	Min	Max
Periodic snapshot data					
Length of the queue (Q)					
Weekday	3,671	3.76	3.81	0	26
Weekend	1,465	6.42	4.90	0	27
Number of employees (E)					
Weekday	3,671	2.11	1.26	0	7
Weekend	1,465	2.84	1.46	0	9
Point-of-sales data					
Purchase incidence of deli products	284,709	22.5%			
Loyalty card data					
Number of visits per customer	13,103	21.7	18.6	7	162

individual customer purchases. Although this sample selection limits the generalizability of our findings, we believe this limitation is not too critical because loyalty card customers are perceived as the most profitable customers by the store. To better control for customer heterogeneity, we focus on grocery purchases of loyalty card customers who visit the store one or more times per month on average. This accounts for a total of 284,709 transactions from 13,103 customers. Table 1 provides some summary statistics describing the queue snapshots and the POS and loyalty card data.

3.2. Purchase Incidence Model

Recall that the POS and loyalty card data are used to construct a panel of observations for each individual customer. Each customer is indexed by i, and each store visit by v. Let $y_{iv} = 1$ if the customer purchased a deli product in that visit, and zero otherwise. Denote $\tilde{Q}_{iv}$ and $\tilde{E}_{iv}$ as the number of people in queue and the number of employees, respectively, that were observed by the customer during visit v. Throughout this paper we refer to $\tilde{Q}_{iv}$ and $\tilde{E}_{iv}$ altogether as the *state of the queue*. The objective of the purchase incidence model is to estimate how the state of the queue affects the probability of purchase of products sold in the deli. Note that we (the researchers) do not observe the state of the queue directly in the data, which complicates the estimation. Our approach is to infer the distribution of the state of queue using snapshot and transaction data and then plug estimates of $\tilde{Q}_{iv}$ and $\tilde{E}_{iv}$ into a purchase incidence model. This methodology is summarized in Figure 2. In this subsection, we describe the purchase incidence model assuming the state of the queue estimates are given (Step 1 in Figure 2); later, §3.3 describes how to handle the unobserved state of the queue.

In the purchase incidence model, the probability of a deli purchase, defined as $p(\tilde{Q}_{iv}, \tilde{E}_{iv}) \equiv \Pr[y_{iv} = 1 \mid \tilde{Q}_{iv}, \tilde{E}_{iv}]$, is modeled as

$$h(p(\tilde{Q}_{iv}, \tilde{E}_{iv})) = f(\tilde{Q}_{iv}, \tilde{E}_{iv}, \beta_q) + \beta_x X_{iv}, \qquad (1)$$

Figure 2 Outline of the Estimation Procedure

- *Step* 0.

(a) Calculate the average store traffic Λ_t using all cashier transactions (including those without deli purchases) for different hours of the day and days of the week (e.g., Mondays between 9 A.M. and 11 A.M.).

(b) Initialize the state of the queue $(\tilde{Q}_{iv}, \tilde{E}_{iv})$ observed by customer i in visit v as the second previous snapshot before checkout time.

(c) Group the snapshot data into *time buckets* with observations for the same time of the day, day of the week and the same number of employees. For example, one bucket could contain snapshots taken on Mondays between 9 A.M. and 11 A.M. with two employees attending. For each time bucket, compute the empirical distribution of the queue length based on the snapshot data.

- *Step* 1. Estimate purchase incidence model (1) via ML assuming state of queue $(\tilde{Q}_{iv}, \tilde{E}_{iv})$ is observed.
- *Step* 2. Estimate the queue intensity ρ_t on each time bucket.

(a) Based on the estimated store traffic Λ_t and purchase incidence probability $p(Q, E)$, calculate the effective arrival rate $\lambda_t(Q, E) = \Lambda_t p(Q, E)$ for each possible state of the queue in time bucket t.

(b) Compute the stationary distribution of the queue length on each time bucket t as a function of the queue intensity ρ_t and $\lambda_t(Q, E)$: for each time bucket, choose the queue intensity ρ_t that best matches the predicted stationary distribution to the observed empirical distribution of the length of the queue (computed in Step 0(c)).

- *Step* 3. Update the distribution of the observed queue length $\tilde{Q}_{iv}$.

(a) Compute the transition probability matrix $P_t(s)$.

(b) For a given deli visit time τ, calculate the distribution of $\tilde{Q}_\tau$ using $P_t(s)$.

(c) Integrate over all possible deli visit times τ to find the distribution of $\tilde{Q}_{iv}$. Update $\tilde{Q}_{iv}$ by its expectation based on this distribution.

(d) Repeat from Step 1 until the estimated length of the queue, $\tilde{Q}_{iv}$, converges.

where $h(\cdot)$ is a link function, $f(\tilde{Q}_{iv}, \tilde{E}_{iv}, \beta_q)$ is a parametric function that captures the impact of the state of the queue, β_q is a parameter vector to be estimated, and X_{iv} is a set of covariates that capture other factors that affect purchase incidence (including an intercept). We use a logit link function, $h(x) = \ln[x/(1-x)]$, which leads to a logistic regression model that can be estimated via maximum likelihood (ML) methods. We tested alternative link functions and found the results to be similar.

Now we turn to the specification of the effect of the state of the queue, $f(\tilde{Q}_{iv}, \tilde{E}_{iv}, \beta_q)$. Previous work has documented that customer behavior is affected by perceptions of waiting that may not be equal to the expected waiting time. Upon observing the state of the queue $(\tilde{Q}_{iv}, \tilde{E}_{iv})$, the measure $W_{iv} = \tilde{Q}_{iv}/\tilde{E}_{iv}$ (number of customers in line divided by the number of servers) is proportional to the expected time to wait in line, and hence is an objective measure of waiting. Throughout this paper, we use the term *expected waiting time* to refer to the *objective* average waiting time faced by customers for a given state of the queue,

which can be different from the *perceived* waiting time they form based on the observed state of the queue. Our first specification uses W_{iv} to measure the effect of this objective waiting factor on customer behavior.

Note that the function $f(W_{iv}, \beta_q)$ captures the *overall* effect of expected waiting time on customer behavior, which includes the disutility of waiting but also potential herding effects. The disutility of waiting has a negative effect, whereas the herding effect has a positive effect. Because both effects occur simultaneously, the estimated overall effect is the sum of both. Hence, the sign of the estimated effect can be used to test which effect dominates. Moreover, as suggested by Larson (1987), the perceived disutility from waiting may be nonlinear. This implies that $f(W_{iv}, \beta_q)$ may not be monotone—herding effects could dominate in some regions whereas waiting disutility could dominate in other regions. To account for this, we specify $f(W_{iv}, \beta_q)$ in a flexible manner using piecewise linear and quadratic functions.

We also estimate other specifications to test for alternative effects. As shown in some of the experimental results reported in Carmon (1991), customers may use the length of the line, $\tilde{Q}_{iv}$, as a visible cue to assess their waiting time, ignoring the *speed* at which the queue moves. In the setting of our pilot study, the length of the queue is highly visible, whereas determining the number of employees attending is not always straightforward. Hence, it is possible for a customer to balk from the queue based on the observed length of the line, without fully accounting for the speed at which the line moves. To test for this, we consider specifications where the effect of the state of the queue is only a function of the queue length, $f(\tilde{Q}_{iv}, \beta_q)$. As before, we use a flexible specification that allows for nonlinear and nonmonotone effects.

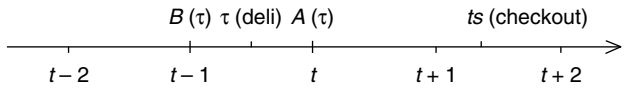
The two aforementioned models look at extreme cases where the state of the queue is fully captured either by the objective expected time to wait (W_{iv}) or by the length of the queue (ignoring the speed of service). These two extreme cases are interesting because there is prior work suggesting each of them as the relevant driver of customer behavior. In addition, $f(\tilde{Q}_{iv}, \tilde{E}_{iv}, \beta_q)$ could also be specified by placing separate weights on the length of the queue ($\tilde{Q}_{iv}$) and the capacity ($\tilde{E}_{iv}$); we also consider these additional specifications in §4.

There are two important challenges to estimate the model in Equation (1). The first is that we are seeking to estimate a causal effect—the impact of $(\tilde{Q}_{iv}, \tilde{E}_{iv})$ on purchase incidence—using observational data rather than a controlled experiment. In an ideal experiment a customer would be exposed to multiple $(\tilde{Q}_{iv}, \tilde{E}_{iv})$ conditions holding all other factors (e.g., prices, time of the day, seasonality) constant. For each of these conditions, her purchasing behavior would then be recorded. In the context of our pilot study, however, there is only one $(\tilde{Q}_{iv}, \tilde{E}_{iv})$ observation for each customer visit. This could be problematic if, for example, customers with a high purchase intention visit the store around the same time. These visits would then exhibit long queues and high purchase probability, generating a bias in the estimation of the causal effect. In fact, the data do suggest such an effect: the average purchase probability is 34.2% on weekends at 8 P.M., when the average queue length is 10.3, and it drops to 28.3% on weekdays at 4 P.M., when the average queue length is only 2.2. Another example of this potential bias is when the deli runs promotions: price discounts attract more customers, which increases purchase incidence and also generates higher congestion levels.

To partially overcome this challenge, we include covariates in X that control for customer heterogeneity. A flexible way to control for this heterogeneity is to include customer fixed effects to account for each customer's average purchase incidence. Purchase incidence could also exhibit seasonality—for example, consumption of fresh deli products could be higher during a Sunday morning in preparation for a family gathering during Sunday lunch. To control for seasonality, the model includes a set of time-of-day dummies interacted with weekend/weekday indicators. This set of dummies also helps to control for a potential endogeneity in the staffing of the deli, because it controls for planned changes in the staffing schedule. Finally, we also include a set of dummies for each day in the sample, which controls for seasonality, trends, and promotional activities (because promotions typically last at least a full day).

Although customer fixed effects account for purchase incidence heterogeneity across customers, they don't control for heterogeneity in purchase incidence across visits of the same customer. Furthermore, some of this heterogeneity across visits may be customer specific, so that they are not fully controlled by the seasonal dummies in the model. State-dependent factors, which are frequently used in the marketing literature (Neslin and van Heerde 2008), could help to partially control for this heterogeneity. Another limitation of the purchase incidence model is that (1) cannot be used to characterize substitution effects with products sold in the prepackaged section, which could be important to measure the overall effect of queue-related factors on total store revenue and profit. To address these limitations, we develop the choice model described in §3.5. Nevertheless, these additions require focusing on a single product category, whereas the purchase incidence model captures all product categories sold in the deli. For this reason and because of its relative simplicity, the estimation of the purchase incidence model (1) provides valuable

Figure 3 **Sequence of Events Related to a Customer Purchase Transaction**



insights about how consumers react to different levels of service.

A second challenge in the estimation of (1) is that $(\tilde{Q}_{iv}, \tilde{E}_{iv})$ are not directly observable in our data set. The next subsection provides a methodology to infer $(\tilde{Q}_{iv}, \tilde{E}_{iv})$ based on the periodic data captured by the snapshots (Q_t, E_t) and describes how to incorporate these inferences into the estimation procedure.

3.3. Inferring Queues from Periodic Data

We start by defining some notation regarding event times, as summarized in Figure 3. Time ts denotes the observed checkout time stamp of the customer transaction. Time $\tau < ts$ is the time at which the customer observed the deli queue and made her decision on whether to join the line (whereas in reality customers could revisit the deli during the same visit hoping to see a shorter line, we assume a single deli visit to keep the econometric model tractable; see Footnote 8 for further discussion). The snapshot data of the queue were collected periodically, generating time intervals $[t-1, t)$, $[t, t+1)$, etc. For example, if the checkout time ts falls in the interval $[t, t+1)$, τ could fall in the intervals $[t-1, t)$, $[t, t+1)$, or in any other interval before ts (but not after). Let $B(\tau)$ and $A(\tau)$ denote the index of the snapshots just before and after time τ. In our application, τ is not observed, and we model it as a random variable and denote $F(\tau \mid ts)$ its conditional distribution given the checkout time ts.[2]

In addition, the state of the queue is only observed at prespecified time epochs, so even if the deli visit time τ is known, the state of the queue is still not known exactly. It is then necessary to estimate (Q_τ, E_τ) for any given τ based on the observed snapshot data (Q_t, E_t). The snapshot data reveal that the number of employees in the system, E_t, is more stable: for about 60% of the snapshots, consecutive observations of E_t are identical. When they change, it is typically by one unit (81% of the samples).[3] When $E_{t-1} = E_t = c$, it seems reasonable to assume that the number of employees remained to be c in the interval $[t-1, t)$. When changes between two consecutive snapshots E_{t-1} and E_t are observed, we assume (for simplicity) that the number of employees is equal to E_{t-1} throughout the interval $[t-1, t)$.

Assumption 1. *In any interval* $[t-1, t)$, *the number of servers in the queuing system is equal to* E_{t-1}.

A natural approach to estimate Q_τ would be to take a weighted average of the snapshots around time τ, for example, an average of $Q_{B(\tau)}$ and $Q_{A(\tau)}$. However, this naive approach may generate biased estimates, as we will show in §3.4. In what follows, we show a formal approach to using the snapshot data in the vicinity of τ to get a point estimate of $\tilde{Q}_\tau$. Our methodology requires the following additional assumption about the evolution of the queuing system:

Assumption 2. *In any snapshot interval* $[t, t+1)$, *arrivals follow a Poisson process with an effective arrival rate* $\lambda_t(Q, E)$ (*after accounting for balking*) *that may depend on the number of customers in queue and the number of servers. The service times of each server follow an exponential distribution with similar rate but independent across servers.*

Assumptions (1) and (2) together imply that in any interval between two snapshots the queuing system behaves like an Erlang queue model (also known as $M/M/c$) with balking rate that depends on the state of queue. The Markovian property implies that the conditional distribution of $\tilde{Q}_\tau$ given the snapshot data only depends on the most recent queue observation before time τ, $Q_{B(\tau)}$, which simplifies the estimation. We now provide some empirical evidence to validate these assumptions.

Given that the snapshot intervals are relatively short (30 minutes), stationary Poisson arrivals within each time interval seem a reasonable assumption. To corroborate this, we analyzed the number of cashier transactions on every half-hour interval by comparing the fit of a Poisson regression model with a negative binomial (NB) regression. The NB model is a mixture model that nests the Poisson model but is more flexible, allowing for overdispersion—that is, a variance larger than the mean. This analysis suggests that there is a small overdispersion in the arrival counts, so that the Poisson model provides a reasonable fit to the data.[4]

The effective arrival rate during each time period $\lambda_t(Q, E)$ is modeled as $\lambda_t(Q, E) = \Lambda_t \cdot p(Q, E)$, where Λ_t is the overall store traffic that captures seasonality

[2] Note that in applications where the time of joining the queue is observed—for example, as provided by a ticket time stamp in a ticket queue—it may still be unobserved for customers that decided not to join the queue. In those cases, τ may also be modeled as a random variable for customers that did not join the queue.

[3] However, there is still sufficient variance of E_t to estimate the effect of this variable with precision; a regression of E_t on dummies for day and hour of the day has an R^2 equal to 0.44.

[4] The NB model assumes Poisson arrivals with a rate λ that is drawn from a gamma distribution. The variance of λ is a parameter estimated from the data; when this variance is close to zero, the NB model is equivalent to a Poisson process. The estimates of the NB model imply a coefficient of variation for λ equal to 17%, which is relatively low.

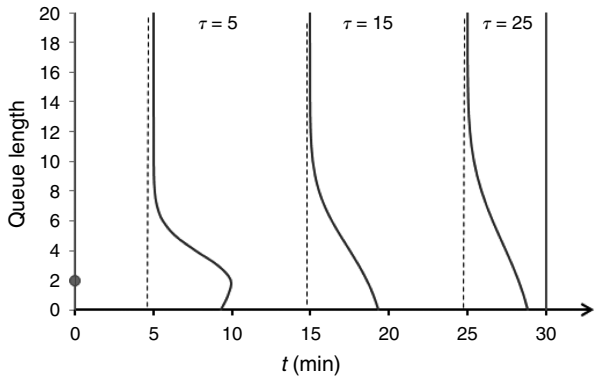
and variations across times of the day; $p(Q, E)$ is the purchase incidence probability defined in (1). To estimate Λ_t, we first group the time intervals into different days of the week and hours of the day and calculate the average number of total transactions in each group, including those without deli purchases (see Step 0(a) in Figure 2). For example, we calculate the average number of customer arrivals across all time periods corresponding to "Mondays between 9 A.M. and 11 A.M." and use this as an estimate of Λ_t for those periods. The purchase probability function $p(Q, E)$ is also unknown; in fact, it is exactly what the purchase incidence model (1) seeks to estimate. To make the estimation feasible, we use an initial rough estimate of $p(Q, E)$ by estimating model (1) replacing $\tilde{E}_\tau$ by $E_{B(ts)-1}$ and $\tilde{Q}_\tau$ by $Q_{B(ts)-1}$ (Step 0(b) in Figure 2). We later show how this estimate is refined iteratively.

Provided an estimate of $\lambda_t(Q, E)$ (Step 2(a) in Figure 2), the only unknown primitive of the Erlang model is the service rate μ_t, or alternatively, the queue intensity level $\rho_t = (\max_Q[\lambda_t(Q, E)])/(E_t \cdot \mu_t)$. Neither μ_t nor ρ_t are observed, and have to be estimated from the data. To estimate ρ_t and also to further validate Assumption 2, we compared the distribution of the observed samples of Q_t in the snapshot data with the stationary distribution predicted by the Erlang model. To do this, we first group the time intervals into *buckets* $\{C_k\}_{k=1}^K$, such that intervals in the same bucket k have the same number of servers E_k (see Step 0(c) in Figure 2). For example, one of these buckets corresponds to "Mondays between 9 A.M. and 11 A.M., with two servers." Using the snapshots on each time bucket, we can compute the observed empirical distribution of the queue. The idea is then to estimate a utilization level ρ_k for each bucket so that the predicted stationary distribution implied by the Erlang model best matches the empirical queue distribution (Step 2(b) in Figure 2). In our analysis, we estimated ρ_k by minimizing the L_2 distance between the empirical distribution of the queue length and the predicted Erlang distribution.

Overall, the Erlang model provides a good fit for most of the buckets: a chi-square goodness of fit test rejects the Erlang model only in 4 out of 61 buckets (at a 5% confidence level). By adjusting the utilization parameter ρ, the Erlang model is able to capture shifts and changes in the shape of the empirical distribution across different buckets. The implied estimates of the service rate suggest an average service time of 1.31 minutes, and the variation across hours and days of the week is relatively small (the coefficient of variation of the average service time is approximately 0.18).[5]

[5] We find that this service rate has a negative correlation (−0.46) with the average queue length, suggesting that servers speed up when the queue is longer (Kc and Terwiesch 2009 found a similar effect in the context of a healthcare delivery service).

Figure 4 **Estimates of the Distribution of the Queue Length Observed by a Customer for Different Deli Visit Times** (τ)



Note. The previous snapshot is at $t = 0$ and shows two customers in queue.

Now we discuss how the estimate of $\tilde{Q}_{i\bar{v}}$ is refined (Step 3 in Figure 2). The Markovian property (given by Assumptions 1 and 2) implies that the distribution of $\tilde{Q}_\tau$ conditional on a prior snapshot taken at time $t < \tau$ is independent of all other snapshots taken prior to t. Given the primitives of the Erlang model, we can use the transient behavior of the queue to estimate the distribution of $\tilde{Q}_\tau$. The length of the queue can be modeled as a birth–death process in continuous time, with transition rates determined by the primitives E_t, $\lambda_t(Q, E)$, and ρ_t. Note that we already showed how to estimate these primitives. The transition rate matrix during time interval $[t, t+1)$, denoted $\mathbf{R_t}$, is given by $[\mathbf{R_t}]_{i,\,i+1} = \lambda_t(i, E_t)$, $[\mathbf{R_t}]_{i,\,i-1} = \min\{i, E_t\} \cdot \mu_t$, $[\mathbf{R_t}]_{i,\,i} = -\sum_{j \neq i}[\mathbf{R_t}]_{i,\,j}$, and zero for the rest of the entries.

The transition rate matrix $\mathbf{R_t}$ can be used to calculate the transition probability matrix for any elapsed time s, denoted $\mathbf{P_t}(s)$.[6] For any deli visit time τ, the distribution of $\tilde{Q}_\tau$ conditional on any previous snapshot $Q_t(t < \tau)$ can be calculated as $\Pr(\tilde{Q}_\tau = k \mid Q_t) = [\mathbf{P_t}(\tau - t)]_{Q_t k}$ for all $k \geq 0$.[7]

Figure 4 illustrates some estimates of the distribution of $\tilde{Q}_\tau$ for different values of τ. (For display purposes, the figure shows a continuous distribution but in practice it is a discrete distribution.) In this example, the snapshot information indicates that $Q_t = 2$, the arrival rate is $\Lambda_t = 1.2$ arrivals/minute and the utilization rate is $\rho = 80\%$. For $\tau = 5$ minutes after the first snapshot, the distribution is concentrated around $Q_t = 2$, whereas for $\tau = 25$ minutes after, the

[6] Using the Kolmogorov forward equations, one can show that $\mathbf{P_t}(s) = e^{\mathbf{R_t}s}$. See Kulkarni (1995) for further details on obtaining a transition matrix from a transition rate matrix.

[7] It is tempting to also use the snapshot after τ, $A(\tau)$, to estimate the distribution of Q_τ. Note, however, that $Q_{A(\tau)}$ depends on whether the customer joined the queue or not, and is therefore endogenous. Simulation studies in §3.4 show that using $Q_{A(\tau)}$ in the estimation of $\tilde{Q}_\tau$ can lead to biased estimates.

distribution is flatter and is closer to the steady state queue distribution. The proposed methodology provides a rigorous approach, based on queuing theory and the periodic snapshot information, to estimate the distribution of the unobserved data $\tilde{Q}_\tau$ at any point in time.

In our application where τ is not observed, it is necessary to integrate over all possible values of τ to obtain the posterior distribution of $\tilde{Q}_{iv}$, so that $\Pr(\tilde{Q}_{iv} = k \mid ts_{iv}) = \int_\tau \Pr(Q_\tau = k)\, dF(\tau \mid ts_{iv})$, where ts_{iv} is the observed checkout time of the customer transaction. Therefore, given a distribution for τ, $F(\tau \mid ts_{iv})$, we can compute the distribution of $\tilde{Q}_{iv}$, which can then be used in Equation (1) for model estimation. In particular, the unobserved value $\tilde{Q}_{iv}$ can be replaced by the point estimate that minimizes the mean square prediction error, i.e., its expected value $E[\tilde{Q}_{iv}]$ (Step 3(b) in Figure 2).[8]

In our application, we discretize the support of τ so that each 30-minute snapshot interval is divided into a grid of 1-minute increments, and calculate the queue distribution accordingly. However, because we do not have precise data to determine the distribution of the elapsed time between a deli visit and the cashier time stamp, an indirect method (described in the appendix) is used to estimate this distribution based on estimates of the duration of store visits and the location of the deli within the supermarket. Based on this analysis, we determined that a uniform in range $[0, 30]$ minutes prior to checkout time is a reasonable distribution for τ.

Assumption 3. *Customers visit the deli once, and this visiting time is uniformly distributed with range* $[0, 30]$ *minutes before checkout time.*

To avoid problems of endogeneity, we determine the distribution of $\tilde{Q}_{iv}$ conditioning on a snapshot that is at least 30 minutes before checkout time (that is, the second snapshot before checkout time) to ensure that we are using a snapshot that occurs before the deli visit time.

Finally, Steps 1–3 in Figure 2 are run iteratively to refine the estimates of effective arrival rate $\lambda_t(Q, E)$, the system intensity ρ_k, and the queue length $\tilde{Q}_{iv}$. In our application, we find that the estimates converge quickly after three iterations.[9]

3.4. Simulation Test

Our estimation procedure has several sources of missing data that need to be inferred: time at which the customer arrives at the deli is inferred from her checkout time, and the state of the queue observed by a customer is estimated from the snapshot data. This subsection describes experiments using simulated data to test whether the proposed methodology can indeed recover the underlying model parameters under Assumptions 1 and 2.

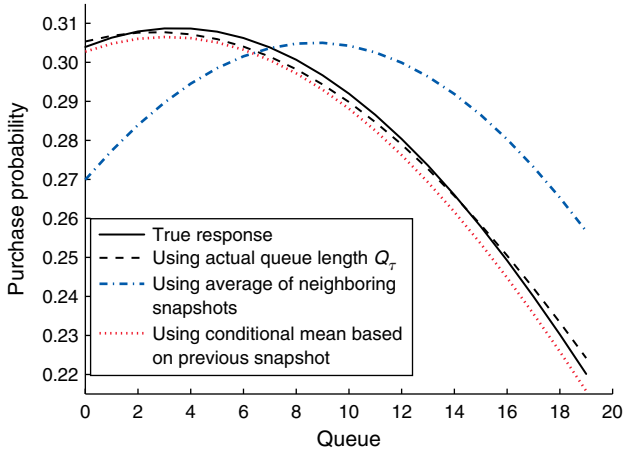
The simulated data are generated as follows. First, we simulate a Markov queuing process with a single server: customers arrive following a Poisson process and join the queue with probability $logit(f(Q))$, where $f(Q)$ is quadratic in Q and has the same shape as we obtained from the empirical purchase incidence model. (We also considered piecewise linear specifications, and the effectiveness of the method was similar.) After visiting the queue, the customer spends some additional random time in the store (which follows a uniform $[0, 30]$ minutes) and checks out. Snapshots are taken to record the queue length every 30 minutes. The arrival rate and traffic intensity are set to be equal to the empirical average value.

Figure 5 shows a comparison of different estimation approaches. The black line, labeled *True response*, represents the customer's purchase probabilities that were used to simulate the data. A consistent estimation should generate estimates that are close to this line. Three estimation approaches, shown with dashed lines in the figure, were compared:

(i) *Using the true state of the queue*, Q_τ. Although this information is unknown in our data, we use it as a benchmark to compare with the other methods. As expected, the purchase probability is estimated accurately with this method, as shown in the black dashed line.

(ii) *Using the average of the neighboring snapshots* $\frac{1}{2}(Q_{B(\tau)} + Q_{A(\tau)})$ *and integrating over all possible values of* τ. Although the average of neighboring snapshots provides an intuitive estimate of Q_τ, this method gives

Figure 5 Estimation Results of the Purchase Incidence Model Using Simulated Data



[8] Although formally the model assumes a single visit to the deli, the estimation is actually using a weighted average of many possible visit times to the deli. This makes the estimation more robust if in reality customers revisit the queue more than once in the hope of facing a shorter queue.

[9] As a convergence criteria, we used a relative difference of 0.1% or less between two successive steps.

biased estimates of the effect of the state of the queue on purchase incidence (the blue dotted-dashed line). This is because $Q_{A(\tau)}$, the queue length in the snapshot following τ, depends on whether the customer purchased or not, and therefore is endogenous (if the customer joins the queue, then the queue following her purchase is likely to be longer). The bias appears to be more pronounced when the queue is short, producing a (biased) positive slope for small values of Q_τ.

(iii) *Using the inference method described in §3.3 to estimate Q_τ, depicted by the red dotted line*. This gives an accurate estimate of the true curve. We conducted more tests using different specifications for the effect of the state of the queue and the effectiveness of the estimation method was similar.

3.5. Choice Model

There are three important limitations of using the purchase incidence model (1). The first is that it does not account for changes in a customer's purchase probability over time, other than through seasonality variables. This could be troublesome if customers plan their purchases ahead of time, as we illustrate with the following example. A customer who does weekly shopping on Saturdays and is planning to buy ham at the deli section visits the store early in the morning when the deli is less crowded. This customer visits the store again on Sunday to make a few "fill-in" purchases at a busy time for the deli and does not buy any ham products at the deli because she purchased ham products the day before. In the purchase incidence model, controls are indeed included to capture the *average* purchase probability at the deli for this customer. However, these controls don't capture the *changes* to this purchase probability between the Saturday and Sunday visits. Therefore, the model would mistakenly attribute the lower purchase incidence on the Sunday visit to the higher congestion at the deli, whereas in reality the customer would not have purchased regardless of the level of congestion at the deli on that visit.

A second limitation of the purchase incidence model (1) is that it cannot be used to attach an economic value to the disutility of waiting by customers. One possible approach would be to calculate an equivalent price reduction that would compensate the disutility generated by a marginal increase in waiting. Model (1) cannot be used for this purpose because it does not provide a measure of price sensitivity. A third limitation is that model (1) does not explicitly capture substitution with products that do not require waiting (e.g., the prepackaged section), which can be useful to quantify the overall impact of waiting on store revenues and profit.

To overcome these limitations, we use a random utility model (RUM) to explain customer choice.

Table 2 Statistics for the 10 Most Popular Ham Products, as Measured by the Percentage of Transactions in the Category Accounted by the Product (Share)

Product	Avg. price	Std. dev. price	Share (%)
1	0.67	0.10	21.23
2	0.40	0.04	9.37
3	0.53	0.06	7.12
4	0.59	0.06	6.13
5	0.64	0.07	5.66
6	0.24	0.01	5.49
7	0.52	0.07	3.97
8	0.54	0.07	3.10
9	0.56	0.07	2.85
10	0.54	0.08	2.20

Note. Prices are measured in local currency per kilogram (one unit of local currency equals approximately US$21).

Because it is common in this type of model, the utility of a customer i for product j during a visit v, denoted U_{ijv}, is modeled as a function of product attributes and parameters that we seek to estimate. Researchers in marketing and economics have estimated RUM specifications using scanner data from a single product category (e.g., Guadagni and Little 1983 model choices of ground coffee products; Bucklin and Lattin 1991 model saltine crackers purchases; Fader and Hardie 1996 model fabric softener choices; Rossi et al. 1996 model choices among tuna products). Note that although deli purchases include multiple product categories, using a RUM to model customer choice requires us to select a single product category for which purchase decisions are independent from choices in other categories and where customers typically choose to purchase at most one SKU in the category. The ham category appears to meet these criteria. The correlations between purchases of ham and other cold-cut categories are relatively small (all less than 8% in magnitude). About 93% of the transactions with ham purchases included only one ham SKU. In addition, it is the most popular category among cold-cuts, accounting for more than 33% of the total sales. The ham category has 75 SKUs, 38 of which are sold in the deli and the rest in the prepackaged section, and about 85% of ham sales are generated in the deli section. In what follows, we describe an RUM framework to model choices among products in the ham category. Table 2 shows statistics for a selection of products in the ham category.

One advantage of using a RUM to characterize choices among SKUs in a category is that it allows us to include product specific factors that affect substitution patterns. Although many of the product characteristics do not change over time and can be controlled by a SKU-specific dummy, our data reveal that prices do fluctuate over time and could be an important driver of substitution patterns. Accordingly, we incorporate product-specific dummies, α_j,

and product prices for each customer visit ($Price_{vj}$) as factors influencing customers' utility for product j. Including prices in the model also allows us to estimate customer price sensitivity, which we use to put a dollar tag on the cost of waiting.

As in the purchase incidence model (1), it is important to control for customer heterogeneity. Because of the size of the data set, it is computationally challenging to estimate a choice model including fixed effects for each customer. Instead, we control for each customer's average buying propensity by including a covariate measuring the average consumption rate of that customer, denoted CR_i. This consumption rate was estimated using calibration data as done by Bell and Lattin (1998). We also use the methods developed by these authors to estimate customers' inventory of ham products at the time of purchase, based on a customers' prior purchases and their consumption rate of ham products. This measure is constructed at the category level and is denoted by Inv_{iv}.

We use the following notation to specify the RUM. Let J be the set of products in the product category of interest (i.e., ham); J_W is the set of products that are sold at the deli section and, therefore, potentially require the customer to wait, $J_{NW} = J \backslash J_W$ is the set of products sold in the prepackaged section, which require no waiting. Let T_v be a vector of covariates that capture seasonal sales patterns, such as holidays and time trends. Define $Fresh_j$ as a binary variable indicating whether product j is sold in the deli (i.e., $j \in J_W$). Using these definitions, customer i's utility for purchasing product j during store visit v is specified as follows:

$$\begin{aligned} U_{ijv} = {} & \alpha_j + \beta_i^q f(\tilde{Q}_{iv}, \tilde{E}_{iv}) Fresh_j + \beta_i^{Fresh} Fresh_j \\ & + \beta_i^{Price} Price_{jv} + \gamma^{cr} CR_i + \gamma^{inv} Inv_{iv} \\ & + \gamma^T T_v + \varepsilon_{ijv}, \end{aligned} \tag{2}$$

where ε_{ijv} is an error term capturing idiosyncratic preferences of the customer, and $f(\tilde{Q}_{iv}, \tilde{E}_{iv})$ captures the effect of the state of the queue on customers' preference. Note that the indicator function $\mathbf{1}[j \in J_W]$ adds the effect of the queue only to the utility of those products that are sold at the deli section (i.e., $j \in J_W$) and not to products that do not require waiting. As in the purchase incidence model (1), the state of the queue $(\tilde{Q}_{iv}, \tilde{E}_{iv})$ is not perfectly observed, but the method developed in §3.3 can be used to replace these by point estimates.[10] An outside good, denoted by $j = 0$, accounts for the option of not purchasing ham, with utility normalized to $U_{i0v} = \varepsilon_{i0v}$. The inclusion of an outside good in the model enables us to estimate how changes in waiting time affect the total sales of products in this category (i.e., category sales).

Assuming a standard extreme value distribution for ε_{ijv}, the RUM described by Equation (2) becomes a random coefficients multinomial logit. Specifically, the model includes consumer-specific coefficients for *Price* (β_i^{Price}); the dummy variable for products sold in the deli (β_i^{Fresh}), as opposed to products sold in the prepackage section; and some of the coefficients associated with the effect of the queue (β_i^q). These random coefficients are assumed to follow a multivariate normal distribution with mean $\boldsymbol{\theta} = (\theta^{Price}, \theta^{Fresh}, \theta^q)'$ and covariance matrix Ω, which we seek to estimate from the data. Including random coefficients for *Price* and *Fresh* is useful to accommodate more flexible substitution patterns based on these characteristics, overcoming some of the limitations imposed by the independence of irrelevant alternatives of standard multinomial logit models. For example, if customers are more likely to switch between products with similar prices or between products that are sold in the deli (or alternatively, in the prepackaged section), then the inclusion of these random coefficients will enable us to model that behavior. In addition, allowing for covariation between β_i^{Price}, β_i^{Fresh}, and β_i^q provides useful information on how customers' sensitivity to the state of the queue relates to the sensitivity to the other two characteristics.

The estimation of the model parameters is implemented using standard Bayesian methods (see Rossi and Allenby 2003). The goal is to estimate (i) the SKU dummies α_j; (ii) the effects of the consumption rate (γ^{cr}), inventory (γ^{inv}), and seasonality controls (γ^T) on consumer utility; and (iii) the distribution of the price and queue sensitivity parameters, which is governed by $\boldsymbol{\theta}$ and Ω. To implement this estimation, we define prior distributions on each of these parameters of interest: $\alpha_j \sim N(\bar{\alpha}, \sigma_\alpha)$, $\gamma \sim N(\bar{\gamma}, \sigma_\gamma)$, $\theta \sim N(\bar{\theta}, \sigma_\theta)$, and $\Omega \sim$ Inverse Wishart(df, Scale). For estimation, we specify the following parameter values for these prior distributions: $\bar{\alpha} = \bar{\gamma} = \bar{\theta} = 0$, $\sigma_\alpha = \sigma_\gamma = \sigma_\theta = 100$, df $= 3$, and Scale equal to the identity matrix. These choices produce weak priors for parameter estimation. Finally, the estimation is carried out using Markov chain Monte Carlo (MCMC) methods. In particular, each parameter is sampled from its posterior distribution conditioning on the data and all other parameter values (Gibbs sampling). When there is no closed-form expression for these full-conditional distributions, we employ Metropolis–Hastings methods (see Rossi and Allenby 2003). The outcome of this estimation process is a sample of values from the posterior distribution of each parameter. Using these values, a researcher can estimate any relevant statistic of

[10] In our empirical analysis, we also performed a robustness check where instead of replacing the unobserved queue length $\tilde{Q}_{iv}$ by point estimates, we sampled different queue lengths from the estimated distribution of $\tilde{Q}_{iv}$. The results obtained with the two approaches are similar.

the posterior distribution, such as the posterior mean, variance, and quantiles of each parameter.

4. Empirical Results

This section reports the estimates of the purchase incidence model (1) and the choice model (2) using the methodology described in §3.3 to impute the unobserved state of the queue.

4.1. Purchase Incidence Model Results

Table 3 reports a summary of alternative specifications of the purchase incidence model (1). All of the specifications include customer fixed effects (11,487 of them), daily dummies (192 of them), and hour-of-day dummies interacted with weekend/holiday dummies (30 of them). A likelihood ratio test indicates that the daily dummies and hour of the day interacted with weekend/holiday dummies are jointly significant (p-value < 0.0001), and so are the customer fixed effects (p-value < 0.0001).

Different specifications of the state of the queue effect are compared, which differ in terms of (1) the functional form for the queuing effect $f(\tilde{Q}, \tilde{E}, \beta_q)$, including linear, piecewise linear, and quadratic polynomial; and (2) the measure capturing the effect of the state of the queue, including (i) expected time to wait, $\tilde{W} = \tilde{Q}/\tilde{E}$, and (ii) the queue length, $\tilde{Q}$ (we omit the tilde in the table). In particular, Models I–III are linear, quadratic, and piecewise linear (with segments at $(0, 5, 10, 15)$) functions of $\tilde{W}$; Model IV–VI are the corresponding models of $\tilde{Q}$. We discuss other models later in this section. The table reports the number of parameters associated with the queuing effects ($\dim(\beta^q)$), the log-likelihood achieved in the maximum likelihood estimation (MLE), and two additional measures of goodness of fit, the Akaike information criterion (AIC) and Bayesian information criterion (BIC), that are used for model selection.

Using AIC and BIC to rank the models, the specifications with $\tilde{Q}$ as explanatory variables (Models IV–VI) all fit significantly better than the corresponding ones with $\tilde{W}$ (Models I–III), suggesting that purchase incidence appears to be affected more by the length of the queue rather than the speed of the service. A comparison of the estimates of the models based on $\tilde{Q}$ is shown in Table 4 and Figure 6 (which plots the results of Model IV–VI). Considering Models V and VI, which allow for a nonlinear effect of $\tilde{Q}$, the pattern obtained in both models is similar: customers appear to be insensitive to the queue length when it is short, but they balk when experiencing long lines. This impact on purchase incidence can become quite large for queue lengths of 10 customers and more. In fact, our estimation indicates that increasing the queue length from 10 to 15 customers would reduce purchase incidence from 30% to 27%, corresponding to a 10% drop in sales.

The AIC scores in Table 3 also suggest that the more flexible models, V and VI, tend to provide a better fit than the less flexible linear model, IV. The BIC score, which puts a higher penalization for the additional parameters, tends to favor the more parsimonious quadratic models, V, and the linear model, IV. Considering both the AIC and BIC score, we conclude that the quadratic specification on queue length (Model V) provides a good balance of flexibility and parsimony, and hence we use this specification as a base for further study.

To further compare the models including queue length versus expected time to wait, we estimated a specification that includes quadratic polynomials of both measures, $\tilde{Q}$ and $\tilde{W}$. Note that this specification nests Models II and V (but it is not shown in the table). We conducted a likelihood ratio test by comparing log-likelihoods of this unrestricted model with the restricted models II and V. The test shows that the coefficients associated with $\tilde{W}$ are not statistically significant, whereas the coefficients associated with $\tilde{Q}$ are. This provides further support that customers put more weight on the length of the line rather than on the expected waiting time when making purchase incidence decisions.

In addition, we consider the possibility that the measure $\tilde{W} = \tilde{Q}/\tilde{E}$ may not be a good proxy for expected time to wait if the service rate of the attending employees varies over time and customers can anticipate these changes in the service rate. Recall, however, that our analysis in §3.3 estimates separate service rates for different days and hours, and shows that there is small variation across time. Nevertheless, we constructed an alternative proxy of expected

Table 3 Goodness-of-Fit Results on Alternative Specifications of the Purchase Incidence Model (Equation (1))

Model	Function form	Metric	Dim(β^q)	Log-likelihood	AIC	Rank	BIC	Rank
I	Linear	*W*	1	−118,195.3	259,808.6	5	382,023.4	3
II	Quadratic	*W*	2	−118,193.1	259,806.2	4	382,031.5	4
III	Piecewise	*W*	4	−118,192.8	259,809.7	6	382,055.8	6
IV	Linear	*Q*	1	−118,189.5	259,797.0	3	382,011.8	1
V	Quadratic	*Q*	2	−118,185.4	259,790.8	1	382,016.0	2
VI	Piecewise	*Q*	4	−118,184.9	259,793.7	2	382,039.8	5

Table 4 MLE Results for Purchase Incidence Model (Equation (1))

	Variable	Coef.	Std. err.	z
Model IV (linear)	$\tilde{Q}$	−0.0133	0.0024	−5.46
Model V (quadratic)	$\tilde{Q}-5.7$	−0.00646	0.00340	−1.90
	$(\tilde{Q}-5.7)^2$	−0.00166	0.00066	−2.50
Model VI (piecewise)	$\tilde{Q}_{0-5}$	0.0056	0.0079	0.71
	$\tilde{Q}_{5-10}$	−0.0106	0.0042	−3.54
	$\tilde{Q}_{10-15}$	−0.0199	0.0068	−2.92
	$\tilde{Q}_{15+}$	−0.0303	0.0210	−1.44

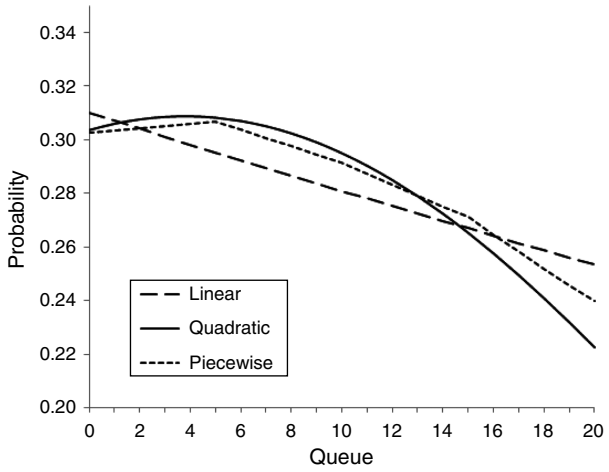
Note. In the quadratic model (Model V), the length of the queue is centered at the mean of 5.7.

time to wait that accounts for changes in the service rate: $W' = \tilde{W}/\mu$, where μ is the estimated service rate for the corresponding time period. Replacing $\tilde{W}$ by W' leads to estimates that are similar to models reported in Table 3.

Although the expected waiting time does not seem to affect customer purchase incidence as much as the queue length, it is possible that customers do take into account the capacity at which the system is operating—i.e., the number of employees—in addition to the length of the line. To test this, we estimated a specification that includes both the queue length $\tilde{Q}$ (as a quadratic polynomial) and the number of servers $\tilde{E}$ as separate covariates.[11] The results suggest that the number of servers, $\tilde{E}$, has a positive impact that is statistically significant, but small in magnitude (the coefficient is 0.0201 with standard error 0.0072). Increasing staff from 1 to 2 at the average queue length only increases the purchase probability by 0.9%. To compare, shortening the queue length from 12 to 6 customers, which is the average length, would increase the purchase probability by 5%. Because both scenarios halve the waiting time, this provides further evidence that customers focus more on the queue length than the objective expected waiting time when making purchase decisions. We also found that the effect of the queue length in this model is almost identical to the one estimated in Model V (which omits the number of servers). We therefore conclude that although the capacity does seem to play a role in customer behavior, its effect is minor relative to the effect of the length of the queue.

Finally, we emphasize that the estimates provide an overall effect of the state of the queue on customer purchases. The estimates suggest that, for queue lengths above the mean (about five customers in line) the effect is significantly negative, which implies that the disutility of waiting seems to dominate any potential herding effects of the queue, whereas for queue lengths below the mean, neither effect is dominant. In our context, herding effects could still be observed, for example, if customers passing by the deli section infer from a long line that the retailer must be offering an attractive deal, or if long lines make the deli section more salient. While the absence of a dominant herding effect seems robust for the *average* customer, we further tested Model V on subsamples of frequent customers (i.e., customers that made 30 or more visits during the study period) and infrequent customers (i.e., customers that made less than 30 visits), with the idea that infrequent customers would be less informed and might potentially learn more from the length of the line. However, we found no significant differences between the estimates. We also partitioned customers into new customers and existing customers (customers are considered to be new within the first two months of their first visit), with the idea that new customers should be less informed.[12] Again, we found no significant differences in the estimated results for the two groups. In summary, the statistical evidence in our results are not conclusive on the presence of dominant herding effects.

Figure 6 Results from Three Different Specifications of the Purchase Incidence Model



4.2. Choice Model Results

In this subsection, we present and discuss the results obtained for the choice model described in §3.5. The specification for the queuing effect $f(\tilde{Q}, \tilde{E})$ is based on the results of the purchase incidence model. In particular, we used a quadratic function of $\tilde{Q}$, which balanced goodness of fit and parsimony in the purchase incidence model. The utility specification includes product-specific intercepts, prices, consumption rate,

[11] We also estimated models with a quadratic term for $\tilde{E}$, but this additional coefficient was not significant.

[12] We used one year of transaction data prior to the study period to verify the first customer visit date. We also tried other definitions of new customers (within three months of the first visit), and the results were similar.

Table 5 Estimation Results for the Choice Model (Equation (2))

	Average effect			Variance/covariance (Ω)	
	Estimate	Std. err.		Estimate	Std. err.
Inv	−0.091	0.026	Ω(*Price*, *Price*)	31.516	1.671
CR	1.975	0.150	Ω(*Fresh*, *Fresh*)	7.719	0.436
Fresh	0.403	0.112	Ω($\tilde{Q}$, $\tilde{Q}$)	0.403	0.083
Price	−9.692	0.203	Ω(*Fresh*, $\tilde{Q}$)	0.020	0.144
$\tilde{Q}$	−0.058	0.061	Ω(*Price*, *Fresh*)	−14.782	0.821
$\tilde{Q}^2$	−0.193	0.122	Ω(*Price*, $\tilde{Q}$)	−0.508	0.267

Note. The estimate and standard error of each parameter correspond to the mean and standard deviation of its posterior distribution, respectively.

household inventory, and controls for seasonality as explanatory variables. The model incorporates heterogeneity through random coefficients for *Price*, the *Fresh* dummy, and the linear term of the length of the queue (Q). We use 2,000 randomly selected customers in our estimation. After running 20,000 MCMC iterations and discarding the first 10,000 iterations, we obtained the results presented in Table 5 (the table omits the estimates of the product-specific intercept and seasonality). The left part of the table shows the estimates of the average effects, with the estimated standard error (measured by the standard deviation of the posterior distribution of each parameter). The right part of the table shows the estimates of the variance–covariance matrix (Ω) characterizing the heterogeneity of the random coefficients β_i^{Price}, β_i^{Fresh}, and $\beta_i^{\tilde{q}}$.

Price, inventory, and consumption rate all have the predicted signs and are estimated precisely. The average of the implied price elasticities of demand is −3. The average effects of the queue coefficients imply qualitatively similar effects as those obtained in the purchase incidence model: consumers are relatively insensitive to changes in the queue length in the $\tilde{Q} = 0$ to $\tilde{Q} = 5$ range, and then the purchase probability starts exhibiting a sharper decrease for queue length values at or above $\tilde{Q} = 6$.

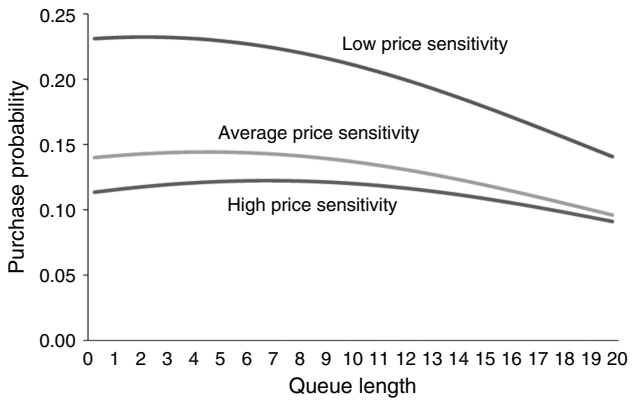
These results can also be used to assign a monetary value to customers' cost of waiting. For example, for an average customer in the sample, an increase from 5 to 10 customers in queue is equivalent to a 1.7% increase in price. Instead, an increase from 10 to 15 customers is equivalent to a 5.5% increase in price, illustrating the strong nonlinear effect of waiting on customer purchasing behavior.

The estimates also suggest substantial heterogeneity in customers' price sensitivities (estimates on the right side of Table 5). The estimated standard deviation of the random price coefficients is 5.614, which implies a coefficient of variation of 57.9%. There is also significant heterogeneity in customer sensitivity to waiting, as measured by the standard deviation of the linear queue effect, which is estimated to be 0.635. The results also show a negative relationship between price and waiting sensitivity and between price and the fresh indicator variable.

To illustrate the implications of the model estimates in terms of customer heterogeneity, we measured the effect of the length of the queue on three customer segments with different levels of price sensitivity: price coefficients equal to the mean, one standard deviation below the mean (labeled high price sensitivity), and one standard deviation above the mean (labeled low price sensitivity), respectively. To compute these choice probabilities, we considered customer visits with average levels of prices, consumption rate, and consumer inventory. Given the negative correlation between the price random coefficient and the two other random coefficients, customers with a weaker price sensitivity will in turn have stronger preferences for fresh products and a higher sensitivity to the length of the queue and, hence, be more willing to wait to buy fresh products. Figure 7 illustrates this pattern, showing a stronger effect of the length of the queue in the purchase probability of the low price sensitivity segment. Interestingly, the low price sensitivity segment is also the most profitable, with a purchase incidence that more than doubles that of the high price sensitivity segment (for small values of the queue length). This has important implications for pricing product categories under congestion effects, as we discuss in the next section.

Finally, because our choice model also considers products that do not require waiting, we measure the extent by which lost sales of fresh products due to a higher queue length are substituted by sales of the prepackaged products. In this regard, our results show that when the length of the line increases, for example, from 5 to 10 customers, only 7% of the deli lost sales are replaced by nondeli purchases. This small substitution effect can be explained by the large heterogeneity of the *Fresh* random coefficient

Figure 7 Purchase Probability for Ham Products in the Deli Section vs. Queue Length for Three Customer Segments with Different Price Sensitivity



together with the relatively small share of purchases of prepackaged products that we observe in the data.

5. Managerial Implications

The results of the previous section suggest that (1) purchase incidence appears to be affected more by the length of the line rather than the speed of the service, and (2) there is heterogeneity in customers' sensitivity to the queue length, which is negatively correlated with their price sensitivity. We discuss three important managerial insights implied by these findings. The first shows that pooling multiple identical queues into a single multiserver queue may lead to an increase in lost sales. The second considers the benefit of adding servers when making staffing decisions. The third discusses the implications of the externalities generated by congestion for pricing and promotion management in a product category.

5.1. Queuing Design

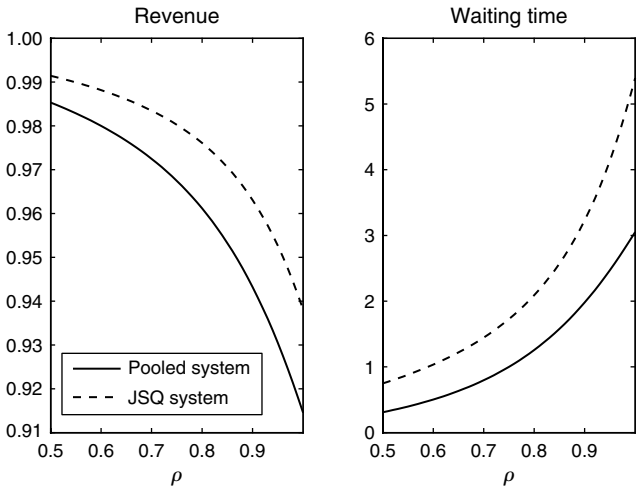
The result from the purchase incidence model that customers react more to the length of the queue than the speed of service has implications on queuing management policies. In particular, we are interested in comparing policies between splitting versus merging queues.

It is well known that an $M/M/c$ pooled queuing system achieves a much lower waiting time than a system with separate $M/M/1$ queues at the same utilization levels. Therefore, if waiting time is the only measure of customer service, then pooling queues is beneficial. However, Rothkopf and Rech (1987) provide several reasons for why pooling queues could be less desirable. For example, there could be gains from server specialization that can be achieved in the separate queue setting. Cachon and Zhang (2007) look at this issue in a setting where two separate queues compete against each other for the allocation of (exogenous) demand and show that using a system with separate queues is more effective (relative to a pooled system) at providing the servers with incentives to increase the service rate. The results in our paper provide another argument for why splitting queues may be beneficial: although the waiting time in the pooled system is shorter, the queue is longer, and this can influence demand. If customers make their decision of joining a queue based on its queue length, as we find in our empirical study, then a pooled system can lead to fewer customers joining the system and therefore increase lost sales. We illustrate this in more detail with the following example.

Consider the following queuing systems: a *pooled* system given by an $M/M/2$ queue with constant arrival rate λ and a *split join-shortest-queue* (*JSQ*) system with two parallel single-server queues with the same overall arrival given by a Poisson process with rate λ and where customers join the shortest queue upon arrival, and assuming that after joining a line customers don't switch to a different line (i.e., no jockeying). If there is no balking—that is, all customers join the queue—it can be shown that the pooled system dominates the split JSQ system in terms of waiting time. However, the queues are longer in the pooled system, so if customers may walk away upon arrival and this balking rate increases with the queue length, then the pooled system may lead to fewer sales.

To evaluate the differences between the two systems, we numerically compute the average waiting time and revenue for both. For the split JSQ system, the approximate model proposed by Rao and Posner (1987) is used to numerically evaluate the system performance. When the queue length is equal to n and the number of servers is c, the arrival rate is $\lambda \Pr(join \mid Q = n, E = c)$, where $\Pr(join \mid Q = n, E = c)$ is customers' purchase probability. In this numerical example, we set the purchase probabilities based on the estimates of the purchase incidence model that includes the quadratic specification of Q. Traffic intensity is defined as $\rho = \max_n \lambda \Pr(join \mid Q = n, E = c)/\mu$, and revenue is defined as the number of customers that join the queue. Figure 8 shows the long-run steady-state average waiting time and average revenue of the two systems. As expected, the pooled $M/M/2$ system always achieves a shorter waiting time. However the $M/M/2$ system generates less revenue because it suffers more traffic loss due to long queues, and the difference increases as the traffic intensity approaches one. In our particular case, the split JSQ system gains 2.7% more revenue while increasing the average waiting time by more than 70% at the highest level of utilization compared to the pooled system. These results imply that when moving toward a pooled system, it may be critical to provide

Figure 8 Comparison Between the Split JSQ and Pooled Systems



information about the expected waiting time so that customers do not anchor their decisions primarily on the length of the line, which tends to increase when the system is pooled.

5.2. Implications for Staffing Decision

The model used in 5.1 also provides insights for making staffing decisions. For example, consider a typical weekday 11:00–12:00 time window versus a weekend 11:00–12:00 window. Given the average customer arrival rates observed at the deli, the minimum capacity needed to meet the demand is one server for the weekday and two for the weekend. The implied utilizations are 75% and 97% for weekdays and weekends, respectively. We use our empirical results to evaluate whether it pays off to add one server in each of these time windows.

In our sample, the average amount that a customer spends at the deli is US$3.3. The estimates from the purchase incidence model suggest that adding a server leads to an increase on purchases of 2% and 7% for the weekday and weekend windows, respectively. This translates into a US$2.3 increase of hourly revenue for the weekday, and a US$20.7 increase for the weekend. In the supermarket of our study, an additional server costs approximately US$3.75 per hour (for full-time staff). The contribution margin is typically in the 10%–25% range for this product category. Hence, it may be profitable to add a server during the weekend 11:00–12:00 period (when the margin is 18% or higher), but not profitable during the weekday 11:00–12:00 period. Interestingly, the supermarket staffing policy seems to be aligned with this result: the snapshot data reveal that between 30%–40% of the time, the deli had a single server staffed during that hour on weekdays, whereas for the weekend more than 75% of the snapshots showed three or more servers.[13]

5.3. Implications for Category Pricing

The empirical results suggest that customers who are more sensitive to prices are less likely to change their probability of purchasing fresh deli products when the length of the queue increases. This can have important implications for the pricing of products under congestion effects, as we show in the following illustrative example.

Consider two vertically differentiated products, H and L, of high and low quality respectively, with respective prices $p_H > p_L$. Customers arrive according to a Poisson process to join an $M/M/1$ queue to buy at most one of these two products. Following model (2), customer preferences are described by an MNL model, where the utility for customer i if buying product $j \in \{L, H\}$ is given by $U_{ij} = \delta_j - \beta_i^p p_j - \beta_i^q Q + \theta_i + \epsilon_{ij}$. Customer may also choose not to join the queue and get a utility equal to $U_{i0} = \epsilon_{i0}$. In this RUM, δ_j denotes the quality of the product, and $\tilde{Q}$ is a random variable representing the queue length observed by the customer upon arrival. Customers have heterogeneous price and waiting sensitivity characterized by the parameters β_i^p and β_i^q. In particular, heterogeneity is modeled through two discrete segments, $s = \{1, 2\}$, with low and high price sensitivity, respectively, and each segment accounts for 50% of the customer population. (Later in this section we will also consider a continuous heterogeneity distribution based on our empirical results.) Let β_1^p and β_2^p be the price coefficients for these segments, with $0 < \beta_1^p < \beta_2^p$. In addition, the waiting sensitivity, β_i^q, is a random coefficient that can take two values: ω_h with probability r_s, and ω_l with probability $1 - r_s$, where s denotes the customer segment and $\omega_l < \omega_h$. This characterization allows for price and waiting sensitivity to be correlated: if $r_1 > r_2$, then a customer with low price sensitivity is more likely to be more waiting-sensitive; if $r_1 = r_2$, then there is no correlation.

Consider first a setting with no congestion so that Q is always zero (for example, if there is ample service capacity). For illustration purposes, we fixed the parameters as follows: $\delta_H = 15$, $p_H = 5$, $\delta_L = 5$, $p_L = 1.5$, $\beta_1^p = 1$, $\theta_1 = 0$, $\beta_2^p = 10$, and $\theta_2 = 12$. In this example, the difference in quality and prices between the two products is sufficiently large so that most of the price sensitive customers ($s = 2$) buy the low quality product L. Moreover, define the cross-price elasticity of demand E_{HL} as the percentage increase in sales of H product from increasing the price of L by 1%, and vice versa for E_{LH}. In this numerical example, we allow for significant heterogeneity with respect to price sensitivity such that, in the absence of congestion, the cross-elasticities between the two products are close to zero (to be exact, $E_{HL} = 0.002$ and $E_{LH} = 0.008$).

Now consider the case where customers observe queues. This generates an externality: increasing the demand of one product generates longer queues, which decreases the utility of some customers who may in turn decide not to purchase. Hence, lowering the price of one product increases congestion and thereby has an indirect effect on the demand of the other product, which we refer to as the *indirect* cross-elasticity effect.

We now show how customer heterogeneity and negative correlation between price and waiting sensitivity can increase the magnitude of the indirect cross-elasticity between the two products. We parametrized

[13] The revenue increase was estimated using specification V from Table 4. We repeated the analysis using a model where customers also account for the number of employees staffed, and the results were similar.

the waiting sensitivity of each segment as $\omega_l = 1.25 - 0.5\Delta$ and $\omega_h = 1.25 + 0.5\Delta$, where Δ is a measure of heterogeneity in waiting sensitivity. We also varied the conditional probabilities r_1 and r_2 to vary the correlation between waiting and price sensitivity while keeping the marginal distribution of waiting sensitivity constant (50% ω_l and 50% ω_h). Fixing all the parameters of the model (including prices p_H and p_L), it is possible to calculate the stationary probabilities of the queue length $\tilde{Q}$. Using the RUM together with this stationary distribution, it is then possible to calculate the share of each product (defined as the fraction of arriving customers that buy each product). Applying finite differences with respect to prices, one can then calculate cross-elasticities that account for the indirect effect through congestion.

Based on this approach, we evaluated the cross-elasticity of the demand for the H product when changing the price of the L product (E_{HL}) for different degrees of heterogeneity in customer sensitivity to wait (Δ) and several correlation patterns. The results of this numerical experiment are presented in Table 6. Note that in the absence of heterogeneity—that is, $\Delta = 0$—the cross-price elasticity is low: the two products H and L appeal to different customer segments, and there is little substitution between them. However, adding heterogeneity and correlation can lead to a different effect. In the presence of heterogeneity, a *negative* correlation between price and waiting sensitivity increases E_{HL}, showing that the *indirect* cross-elasticity increases when the waiting sensitive customers are also the least sensitive to price. The changes in cross-elasticity due to correlation can become quite large for higher degrees of customer heterogeneity. In the example, when $\Delta = 2$, the cross-elasticity changes from 0.011 to 0.735 when moving from positive to negative correlation patterns.

We now discuss the intuition behind the patterns observed in the example of Table 6. When there is heterogeneity in price sensitivity, lowering the price of the L product attracts customers who were not purchasing before the price reduction (as opposed to cannibalizing the sales of the H product). Because of this increase in traffic, congestion in the queue increases, generating longer waiting times for all customers.

Table 6 **Cross-Price Elasticities Describing Changes in the Probability of Purchase of the High-Price Product (H) from Changes in the Price of the Low-Price Product (L)**

	Correlation between price and waiting sensitivity				
Heterogeneity	−0.9	−0.5	0	0.5	0.9
$\Delta = 0.0$	—	—	0.042	—	—
$\Delta = 1.0$	0.342	0.228	0.120	0.047	0.010
$\Delta = 2.0$	0.735	0.447	0.209	0.070	0.011

But when price and waiting sensitivity are negatively correlated, the disutility generated by the congestion will be higher for the less price sensitive customers, and they will be more likely to walk away after the price reduction in L. Because a larger portion of the demand for the H product comes from the less price sensitive buyers, the indirect cross-price elasticity will increase as the correlation between price and waiting sensitivity becomes more negative.

Although the above example uses discrete customer segments, similar effects occur when considering heterogeneity described through a continuous distribution, as in our empirical model. Similar to the previous discrete case example, we assume the utility for customer i to purchase j is given by $U_{ij} = \delta_j - \beta_i^p p_j + f(\beta_i^q, Q) + \theta_i$. But now the queue effect is specified by the quadratic form with random coefficients for $(\beta^p, \beta^q, \theta)$, which are normally distributed with the same covariance matrix as the one estimated in Table 5. Prices p_L and p_H are picked to reflect the true price of high end and low end products, and λ to reflect the empirical average arrival rate in the deli session. In this case, our calculation shows a cross-price elasticity equal to $E_{HL} = 0.81$. In a counterfactual that forces the waiting sensitivity β^q to be independent of the other random coefficients (β^p, θ), the price elasticity E_{HL} drops to 0.083, one order of magnitude smaller, showing qualitatively similar results to those from the discrete heterogeneity example.

In summary, the relationship between price and waiting sensitivity is an important factor affecting the prices in a product category when congestion effects are present. Congestion can induce price–demand interactions among products that in the absence of congestion would have a low direct cross-price-elasticity of demand. Our analysis illustrates how heterogeneity and negative correlation between price and waiting sensitivity can exacerbate these interactions through stronger indirect cross-elasticity effects. This can have important implications on how to set prices in the presence of congestion.

6. Conclusions

In this paper, we use a new data set that links the purchase history of customers in a supermarket with objective service level measures to study how an important component of the service experience—waiting in queue—affects customer purchasing behavior.

An important contribution of this paper is methodological. An existing barrier to studying the impact of service levels on customer buying behavior in retail environments comes from the lack of objective data on waiting time and other customer service metrics. This work uses a novel data collection

technique to gather high frequency store operational metrics related to the actual level of service delivered to customers. Because of the periodic nature of these data, an important challenge arises in linking the store operational data with actual customer transactions. We develop a new econometric approach that relies on queuing theory to infer the level of service associated with each customer transaction. In our view, this methodology could be extended to other contexts where periodic service level metrics and customer transaction data are available. This methodology also enables us to estimate a comprehensive descriptive model of how waiting in queue affects customer purchase decisions. Based on this model, we provide useful prescriptions for the management of queues and other important aspects of service management in retail. In this regard, a contribution of our work is to measure the overall impact of the state of the queue on customer purchase incidence, thereby attaching an economic value to the level of service provided. This value of service together with an estimate of the relevant operating costs can be used to determine an optimal target service level, a useful input for capacity and staffing decisions.

Second, our approach empirically determines the most important factors in a queuing system that influence customer behavior. The results suggest that customers seem to focus primarily on the length of the line when deciding to join a queue, whereas the number of servers attending the queue, which determines the speed at which the queue advances, has a much smaller impact on customers' decisions. This has implications for the design of a queuing system. For example, although there are several benefits of pooling queues, the results in this paper suggest that some precautions should be taken. In moving toward a pooled system, it may be critical to provide information about the expected waiting time so that customers are not drawn away by longer queues. In addition, our empirical analysis provides strong evidence that the effect of waiting on customer purchases is nonlinear. Hence, measuring extremes in the waiting distribution—for example, the fraction of the time that 10 or more customers are waiting in queue—may be more appropriate than using average waiting time to evaluate the system's performance.

Third, our econometric model can be used to segment customers based on their waiting and price sensitivities. The results show that there is indeed a substantial degree of heterogeneity in how customers react to waiting and price, and moreover, the waiting and price sensitivity are negatively correlated. This has important implications for the pricing of a product category where congestion effects are present. Lowering prices for one product increases demand for that alternative, but also raises congestion, generating a negative externality for the demand of other products from that category. Heterogeneity and negative correlation in price and waiting sensitivity exacerbate this externality, and therefore should be accounted for in category pricing decisions. We hope that this empirical finding fosters future analytical work to study further implications of customer heterogeneity on pricing decisions under congestion.

Finally, our study has some limitations that could be explored in future research. For example, our analysis focuses on studying the short-term implications of queues by looking at how customer purchases are affected during a store visit. There could be long-term effects whereby a negative service experience also influences future customer purchases, for example, the frequency of visits and retention. Another possible extension would be to measure how observable customer characteristics—such as demographics—are related to their sensitivity to wait. This would be useful, for example, to prescribe target service levels for a new store based on the demographics of the market. Competition could also be an important aspect to consider; this would probably require data from multiple markets to study how market structure mediates the effect of queues on customer purchases.

On a final note, this study highlights the importance of integrating advanced methodologies from the fields of operations management and marketing. We hope that this work stimulates further research on the interface between these two academic disciplines.

Acknowledgments

The authors thank seminar participants at the University of North Carolina at Chapel Hill, Duke University, Harvard University, the Consortium for Operational Excellence in Retail, the 2011 Marketing Science Conference, the 2011 INFORMS Conference, the 2011 London Business School Innovation in Operations Conference, the 2011 M&SOM Service Operations Special Interest Group, and the 2011 Wharton Workshop on Empirical Research in Operations Management for their useful feedback. They also thank SCOPIX for providing the data used in this study. Marcelo Olivares acknowledges partial support from FONDECYT [Grant 1120898].

Appendix. Determining the Distribution for Deli Visit Time

Our estimation method requires integrating over different possible values of deli visit time. This appendix describes how to obtain an approximation of this distribution. Our approach follows two steps. First, we seek to estimate the distribution of the duration of a supermarket visit. Second, based on the store layout and previous research on customer paths in supermarket stores, we determine (approximately) in which portion of the store visit customers would cross the deli.

In terms of the first step, to get an assessment of the duration of a customer visit to the store, we conducted some

Table A.1 Regression Results

	Estimate	Std. err.
θ_0	0.069*	(0.021)
θ_1	0.107*	(0.026)
θ_2	0.101*	(0.027)
θ_3	0.063*	(0.027)
θ_4	0.066*	(0.027)
θ_5	0.029	(0.027)
θ_6	−0.020	(0.023)
N	879	
R^2	0.928	

Note. Standard errors are in parentheses.
*$p < 0.05$.

additional empirical analysis using store foot traffic data. Specifically, we collected data on the number of customers that entered the store during 15-minute intervals (for the month of February of 2009). With these data, our approach requires discretizing the duration of a visit in 15-minute time intervals. Accordingly, let T denote a random variable representing the duration of visit, from entry until finishing the purchase transaction at the cashier, with support in $\{0, 1, 2, 3, 4, 5, 6\}$; $T = 0$ is a visit of 15 minutes or less, $T = 1$ corresponds to a visit between 15 and 30 minutes, and similarly for the other values. Let $\theta_t = \Pr(T = t)$ denote the probability mass function of this random variable. Not all customers that enter the store go through the cashier: with probability ψ a customer leaves the store without purchasing anything. Hence, $\sum_{t=0}^{6} \theta_t + \psi = 1$. Note that $\{\theta_t\}_{t=0,\ldots,6}$ and ψ completely characterize the distribution of the visit duration T.

Let X_t be the number of entries observed during period t, and let Y_t be the total number of observed transactions in the cashiers during that period. We have

$$E(Y_t \mid \{X_r\}_{r \le t}) = \sum_{s=0}^{6} X_{t-s}\theta_s.$$

Because the conditional expectation of Y_t is linear in the contemporaneous and lagged entries $X_t, \ldots, X_{t-6}$, the distribution of the duration of the visit can be estimated through the linear regression

$$Y_t = \sum_{s=0}^{6} X_{t-s}\theta_s + u_t.$$

Note that the regression does not have an intercept. Table A.1 shows the ordinary least squares (OLS) estimates of this regression.[14]

The parameters θ_0 through θ_4 are positive and statistically significant. (The other parameters are close to zero and insignificant, so we consider those being equal to zero.) Conditional on going through the cashier, approximately 70% of the customers spend 45 minutes or less in the store (calculated as $\sum_{t=0}^{2} \theta_t / \sum_{t=0}^{4} \theta_t$), and 85% of them less than an hour. The average duration of a visit is approximately 35 minutes. Moreover, the distribution of the duration of the store visit could be approximated reasonably well by a uniform distribution with range $[0, 75]$ minutes.

To further understand the time at which a customer visits the deli, it is useful to understand the path that a customer follows during a store visit. In this regard, the study by Larson et al. (2005) provides some information of typical customer shopping paths in supermarket stores. They show that most customers tend to follow a shopping path through the "racetrack"—the outer ring of the store that is common in most supermarket layouts. In fact, the supermarket where we base our study has the deli section located in the middle of the racetrack. Moreover, Hui et al. (2009) show that customers tend to buy products in a sequence that minimizes total travel distance. Hence, if customer baskets are evenly distributed through the racetrack, it is likely that the visit to the deli is done during the middle of the store visit. Given that the visit duration tends to follow a uniform distribution between $[0, 75]$ minutes, we approximate the distribution of deli visit time by a uniform distribution with range $[0, 30]$ minutes before checkout time.

[14] The parameters of the regression could be constrained to be positive and to sum to less than one. However, in the unconstrained OLS estimates, all the parameters that are statistically significant satisfy these constraints.

References

Afanasyev M, Mendelson H (2010) Service provider competition: Delay cost structure, segmentation, and cost advantage. *Manufacturing Service Oper. Management* 12(2):213–235.

Afèche P, Mendelson H (2004) Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Management Sci.* 50(7):869–882.

Akşin Z, Ata B, Emadi SM, Su C-L (2013) Structural estimation of callers' delay sensitivity in call centers. *Management Sci.* Forthcoming.

Allon G, Federgruen A, Pierson M (2011) How much is a reduction of your customers' wait worth? An empirical study of the fast-food drive-thru industry based on structural estimation methods. *Manufacturing Service Oper. Management* 13(4):489–507.

Antonides G, Verhoef PC, van Aalst M (2002) Consumer perception and evaluation of waiting time: A field experiment. *J. Consumer Psych.* 12(3):193–202.

Bell DR, Lattin JM (1998) Shopping behavior and consumer preference for store price format: Why "large basket" shoppers prefer EDLP. *Marketing Sci.* 17(1):66–88.

Berry LL, Seiders K, Grewal D (2002) Understanding service convenience. *J. Marketing* 66(3):1–17.

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

Bucklin RE, Lattin JM (1991) A two-state model of purchase incidence and brand choice. *Marketing Sci.* 10(1):24–39.

Cachon GP, Zhang F (2007) Obtaining fast service in a queueing system via performance-based allocation of demand. *Management Sci.* 53(3):408–420.

Campbell D, Frei F (2011) Market heterogeneity and local capacity decisions in services. *Manufacturing Service Oper. Management* 13(1):2–19.

Carmon Z (1991) Recent studies of time in consumer behavior. *Adv. Consumer Res.* 18(1):703–705.

Chandon P, Morwitz VG, Reinartz WJ (2005) Do intentions really predict behavior? Self-generated validity effects in survey research. *J. Marketing* 69(2):1–14.

Davis MM, Vollmann TE (1993) A framework for relating waiting time and customer satisfaction in a service operation. *J. Services Marketing* 4(1):61–69.

Deacon RT, Sonstelie J (1985) Rationing by waiting and the value of time: Results from a natural experiment. *J. Political Econom.* 93(4):627–647.

Debo L, Veeraraghavan S (2009) Models of herding behavior in operations management. Netessine S, Tang CS, eds. *Consumer-Driven Demand and Operations Management, International Series in Operations Research and Management Science*, Vol. 131, Chap. 4 (Springer Science, New York), 81–111.

Fader PS, Hardie BGS (1996) Modeling consumer choice among SKUs. *J. Marketing Res.* 33(4):442–452.

Fisher ML, Krishnan J, Netessine S (2009) Are your staffing levels correct? *Internat. Commerce Rev.* 8(2):110–115.

Forbes SJ (2008) The effect of air traffic delays on airline prices. *Internat. J. Indust. Organ.* 26(5):1218–1232.

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

Gross D, Shortle JF, Thompson JM, Harris CM (2008) *Fundamentals of Queueing Theory*, 4th ed. (Wiley, Hoboken, NJ).

Guadagni PM, Little JDC (1983) A logit model of brand choice calibrated on scanner data. *Marketing Sci.* 2(3):203–238.

Hasija S, Pinker E, Shumsky R (2008) Call center outsourcing contracts under information assymetry. *Management Sci.* 54(4):793–807.

Hess S, Bierlaire M, Polak JW (2005) Estimation of value of travel-time savings using mixed logit models. *Transportation Res. Part A: Policy Practice* 39(2–3):221–236.

Hui MK, Tse DK (1996) What to tell consumers in waits of different lengths: An integrative model of service evaluation. *J. Marketing* 60(2):81–90.

Hui SK, Fader PS, Bradlow ET (2009) The traveling salesman goes shopping: The systematic deviations of grocery paths from TSP optimality. *Marketing Sci.* 28(3):566–572.

Hui MK, Laurette D, Chebat JC (1997) The impact of music on consumers reactions to waiting for services. *J. Retailing* 73(1):87–104.

Ibrahim R, Whitt W (2011) Wait-time predictors for customer service systems with time-varying demand and capacity. *Oper. Res.* 59(5):1106–1118.

Janakiraman N, Meyer RJ, Hoch SJ (2011) The psychology of decisions to abandon waits for service. *J. Marketing Res.* 48(6): 970–984.

Katz KL, Larson BW, Larson RC (1991) Prescription for the waiting-in-line blues: Enlighten, entertain, and engage. *Sloan Management Rev.* 32(2):44–53.

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

Kulkarni VG (1995) *Modeling and Analysis of Stochastic Systems* (Chapman & Hall/CRC, London).

Larson JS, Bradlow ET, Fader PS (2005) An exploratory look at supermarket shopping paths. *Internat. J. Res. Marketing* 22(4):395–414.

Larson RC (1987) Perspective on queues: Social justice and the psychology of queueing. *Oper. Res.* 35(6):895–905.

Mandelbaum A, Zeltyn S (2004) The impact of customers patience on delay and abandonment: Some empirically-driven experiments with the mmng queue. *OR Spectrum* 26(3):377–411.

Neslin SA, van Heerde HJ (2008) Promotion dynamics. *Foundations Trends Marketing* 3(4):177–268.

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

Png IPL, Reitman D (1994) Service time competition. *RAND J. Econom.* 25(4):619–634.

Rao BM, Posner MJM (1987) Algorithmic and approximation analyses of the shorter queue model. *Naval Res. Logist.* 34(3):381–398.

Rossi PE, Allenby GM (2003) Bayesian statictics and marketing. *Marketing Sci.* 22(3):304–328.

Rossi PE, McCulloch RE, Allenby GM (1996) The value of purchase history data in target marketing. *Marketing Sci.* 15(3):321–240.

Rothkopf MH, Rech P (1987) Perspectives on queues: Combining queues is not always beneficial. *Oper. Res.* 35(6):906–909.

Taylor S (1994) Waiting for service: The relationship between delays and evaluations of service. *J. Marketing* 58(2):56–69.

Train K (2003) *Discrete Choice Methods with Simulation* (Cambridge University Press, Cambridge, UK).