informs
http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–29
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Designing Pricing Strategy for Operational and Technological Transformation

Vineet Kumar,[a] Yacheng Sun[b]

[a] School of Management, Yale University, New Haven, Connecticut 06511; [b] Department of Marketing, Tsinghua University, Beijing 100084, China

Contact: vineet.kumar@yale.edu, http://orcid.org/0000-0001-8784-6858 (VK); sunyc@sem.tsinghua.edu.cn (YS)

Received: June 14, 2015
Revised: February 14, 2017; October 1, 2018
Accepted: December 10, 2018
Published Online in Articles in Advance: ■■ ■■, 2019

https://doi.org/10.1287/mnsc.2019.3315

Abstract. We examine how operational or technological transformation impacts consumer value, as well as the effectiveness of a firm's pricing strategies. We develop a model of multidimensional screening featuring forward-looking consumers who make short-run consumption and long-run purchase decisions. Using a detailed panel of consumer data from a rental-by-mail firm, we estimate consumer utility for current consumption, obtaining heterogeneous preferences for bunching and smoothing consumption. Using counterfactual analysis, we evaluate the impact of improving service time. We find that the firm with improved service time might create more value for all consumers, but its profits and even revenues could diminish because value extraction becomes more difficult. We find a novel mechanism that causes this effect, which is driven by *increased consumer heterogeneity* in the valuation for each product and *reduced differentiation* across products. This result persists even when the firm can reoptimize its price levels based on the service time. We find that a change in the pricing strategy might be required for the firm to obtain higher revenue with improved service time.

History: Accepted by Matthew Shum, marketing.
Funding: Vineet Kumar thanks Yale School of Management for financial support. Yacheng Sun acknowledges the financial support from Tsinghua University's Special Funds for the Independent Research in Arts and Sciences, the Natural Science Foundation of China [Grants 91746302 and 71532006] and the China Ministry of Education Project for Key Research Institute of Humanities and Social Sciences in Universities [Grant 16JJD630006].

Keywords: pricing strategy • operations • dynamic structural model

1. Introduction

The operational and technological environments that firms face often undergo significant changes that impact both firms and consumers. We empirically study how such improvements impact consumers and firms. When such changes improve service quality or service time, they have the potential to create significantly more value for consumers, and they may have differential impacts across multiple dimensions of value. Firms also need to tailor their pricing to continue to be able to extract the surplus that is created in such a case. More specifically, we develop a model from microfoundations to examine the following questions:

a. How do improvements in service time due to operations or technology impact consumers' consumption choices and willingness to pay for the service? How are the different dimensions of consumer value each impacted by the service time improvement?

b. Do operations or technology quality improvements (by decreasing service time) make customer segmentation easier or more difficult? What are the revenue and profit impacts of these quality improvements? Does the firm always benefit from such improvements?

c. In a multidimensional screening setting, which pricing mechanisms allow the firm to capture greater value? Which mechanisms are better able to achieve the trade-off between surplus creation and surplus extraction for long and short service time?

Our empirical setting is that of a firm providing rental-by-mail (RBM) movies—the Netflix DVD model, whereby the service or service time plays a critical role in consumer valuations of the firm's product. We develop a model of consumer choices from microfoundations, to be able to characterize how the service time impacts consumer utilities, and consequently, choices.

The unique data contain a representative sample of consumers, for whom we observe monthly payments, and detailed, daily-level movie shipping records. The data reveal a number of interesting patterns. First, we find that consumers watch many fewer movies than the plan allows, foregoing immediate consumption. Second, consumption is more likely to occur when consumers hold a long level of inventory or have shorter service times, everything else being equal. Third, consumers exhibit inertia in plan choice. Fourth, the service time (operational effectiveness) has a significant effect on consumers' demand dynamics. Each of these findings has important managerial implications.

We develop an empirical framework of consumers' shorter-term (daily) consumption choices and longer-term (monthly) purchase choices, which captures the empirical features of the RBM model. The framework is readily customizable to other settings. This model of quantity choice features an ordered choice framework for consumption and plan choices; it is developed from microfoundations and can be applied to similar choice settings. We use this framework to (a) characterize consumers' dynamic decision-making processes and quantify the distribution of their valuations of the service; and (b) assess the effects of the service time resulting from improved operations or technological changes. We evaluate how firm costs, which are driven by consumers' consumption decisions, and revenue, derived from purchase decisions, are both impacted by changes due to operational and technological changes.

We model the intertemporal trade-offs inherent in consumer dynamics across short-run and long-run time scales. First, in the short-run dynamics, the consumer trades off immediate versus future consumption. Consumers in our model are heterogeneous in multiple dimensions, in terms of their taste for consumption and in terms of whether they prefer intertemporal smoothing or bunching. Second, with regard to long-run dynamics, the consumer trades off the subscription price and the additional flexibility in consumption (e.g., more movies in the mail). This decision is impacted by the set of products available from the firm. Our model explicitly characterizes these two types of trade-offs induced by the plan quota and service time.

Estimating the model with heterogeneous consumers brings about a number of computational challenges. Modeling both the short-run and long-run dynamics requires us to characterize a very large state space. We adapt the estimation framework developed by Imai et al. (2009) (IJC), which allows a flexible, hierarchical Bayesian model of individual-level heterogeneity.

From the firm's perspective, we have a multidimensional screening problem whereby the firm offers a menu of contracts and heterogeneous consumers self-select. We examine a number of counterfactual scenarios by altering the service time. We find a nonmonotonic relationship between operational performance and both profits and even revenues for the firm. We uncover two general mechanisms underlying these findings and find that reducing service time can increase consumer heterogeneity in valuations, making it more difficult to extract surplus. To the best of our knowledge, this effect and mechanism have not been examined or suggested, either empirically or theoretically. Further, we show that the firm could overcome this pitfall by changing its pricing strategy, either by charging a unit price or by customizing the subscription prices according to the service time. Finally, we examine the case when the service time is zero, with this instantaneous delivery motivated by streaming services, such as Netflix. We show that the bundling of the streaming service using monthly subscriptions can be profitable, but only when marginal costs are fairly low. These findings illustrate the importance of understanding the interaction between value creation (based on service time) and value extraction (based on pricing strategy) and show how they must be aligned to avoid a harmful outcome.

More broadly, vertical quality-type improvements are considered to be better for surplus generation for consumers, as well as for surplus capture by firms. We demonstrate the critical role of the pricing strategy in this argument. We show that reoptimizing the price levels to account for such improvements may not be sufficient and that exogenously improving service time ("quality") might be harmful. Specifically, the firm's revenue and profits could decrease with service time improvements, unless the firm changes its pricing strategy or value capture mechanism.

Another way to view our results is that business model choices that were appropriate for low operational effectiveness might prove harmful for profitability when we have operational improvements. We empirically demonstrate how pricing strategy has the ability to shape both how much value is created for consumers and how much value can be captured. Consumer heterogeneity in product valuations plays a critical role and can increase when operations and technology improve, and we are able to characterize the valuation with our microfoundations-based model.

Our major contributions involve the examination of consumer choices, as well as optimal firm pricing strategy under changing operational or technological environments. First, there are few empirical characterizations of how operational effectiveness (service time) impacts consumer value (willingness to pay [WTP]) and different dimensions of value (we examine both consumption value and option value). Specifically, we quantify how firm actions impact the distribution of valuations (WTP) and how these change for each of the firm's product offerings, owing to improvements in service time.

Our results suggest that improvements in service time resulting from operations or technology can make it more challenging for the firm to extract surplus. There are two underlying reasons for this result, and we introduce a new mechanism to the literature at the interface of marketing and operations. First, we find that the heterogeneity in valuations *for each product (plan)* increases when service time is reduced. Increased heterogeneity makes it more challenging for the firm to capture surplus. Second, we find that with improved service time, consumers' valuations *across the products* actually become more similar, making segmentation much more challenging using a menu of product offerings. Both these factors diminish the firm's ability to extract its share of the surplus, even though improved service time unambiguously increases the total available surplus.

Second, we demonstrate how moving from one pricing strategy to another impacts revenues and profitability, incorporating the strategic responses of forward-looking customers. Third, we empirically identify how pricing mechanisms that create the most surplus may be less profitable for the firm. We also characterize the pricing strategy (mechanism) that improves surplus extraction as operational effectiveness increases. Finally, we show that the firm needs to change not only price levels but also its pricing strategy in response to improved operations; otherwise, it might actually generate lower revenue with operational improvements. Overall, we find that the potential misalignment between operational efficiency and the pricing strategy can be costly, in terms of missed profits in the short run, and tempered incentives for the firm to invest in operational and technology changes in the long run. Firms need a systematic framework to understand consumers' consumption and purchase decisions, before and after the operational and technological changes. Consequently, the firm can adapt its pricing strategies accordingly to be aligned with the operational and technology changes and to identify the appropriate boundary conditions for the applicability of alternative pricing options.

The rest of the manuscript is organized as follows. We examine related literature (Section 2) and provide institutional details regarding the setting and data patterns (Section 3). We then present the model (Section 4), identification and estimation (Section 5), results (Section 6), counterfactuals (Section 7), and discussion (Section 8).

2. Literature Review

Our research contributes to multiple streams of literature. First is the link between the operations and marketing capabilities of the firm, which contribute to the overall profitability of the firm. There are few empirical studies that have demonstrated how the effectiveness or pricing or other marketing strategies are impacted by operational or technological transformation. We both demonstrate this connection in our empirical setting and uncover a novel mechanism that causes the effects.

Our paper is closely connected to the price discrimination and product differentiation literature, especially second-degree price discrimination with consumer self-selection from a price–quantity menu (Mussa and Rosen 1978, Maskin and Riley 1984). Because subscription plans vary by the amounts of quotas, subscription pricing is a special form of second-order price discrimination studied under monopoly and competition (Rochet and Stole 2002, Stole 2003, Crawford and Shum 2007). A related study has examined the potential value of first-degree price discrimination in Netflix movie rentals, using high-dimensional data to connect browsing behavior and consumer characteristics to valuations (Shiller 2014). Informed by the *throttling* practice of Netflix, the RBM rental firm could also implement location-based pricing (Miller and Osborne 2014, Ngwe 2017). Firms that offer multiple quantity levels are also known to distort their quantities from the first best under incomplete information (McManus 2007). However, in our case the plan quota is discrete with $Q = 1$, 2, or 3.

However, this stream of literature has not considered the impact of operational or technological changes, which cause the novel mechanism for lower profitability under better service time. We believe that our contribution is the first to empirically make this connection. Through the lens of the multidimensional screening literature (Armstrong and Rochet 1999), consumers of the RBM service may have different value dimensions. Thus, the pricing decision by our focal firm involves using different subscription plans to separate consumers. To the best of our knowledge, we are the first to study the multidimensional screening problem in the context of different pricing strategies.

Another stream of literature evaluates the rental-by-mail business model, with theoretical models. Among the issues examined are the implications for the rental firm's purchasing, stocking, and inventory allocation decisions (Bassamboo et al. 2009), waiting costs (Cachon and Feldman 2011), service time, and usage and pricing strategies (Randhawa and Kumar 2008, Tong and Rajagopalan 2014). However, our primary questions of interest have not been examined here.

More broadly, our empirical framework models individual, daily-level consumption decisions and flexibly captures both the observed drivers of consumption decisions and unobservables (e.g., the amount of time available to watch movies) that may create correlations between consumption decisions. Unlike existing studies, our framework explicitly recognizes the closed-loop delivery process of the RBM process and the temporal interdependence between consumption decisions. Second, these studies also greatly simplify consumers' purchase decisions, so that the firm's pricing decisions are limited to setting a single subscription fee or a per-usage fee. Our distinct framework integrates both consumption and purchase decisions and examines a wide range of pricing strategies.

3. Empirical Setting and Data

Our empirical setting is the rental-by-mail business model. RBM services are widely adopted in the United States: well-known examples include movie rentals (e.g., Netflix), games (e.g., Gamefly), books (e.g., Bookswim), and apparel (e.g., RentTheRunway).

RentTheRunway alone now serves 9 million subscribers.[1] As the name suggests, RBM services typically use the U.S. Postal Service or courier services to deliver the rental products to the consumer. The delivery process is "closed loop"; that is, the company ships a "new" rental product only after it receives a returned product from the consumer. An important operations aspect of the delivery RBM process, detailed in Appendix A, is the service time, or the time it takes for a complete shipment cycle (see Figure A.1). RBM services predominantly use a subscription pricing model, whereby the consumer chooses from a menu of subscription plans for a specified period—typically one month. Each plan is characterized by a price and a mailing quota (number of rental products checked out at a time).

We next describe the data and then examine model-free evidence as well as evidence from reduced-form models. Our objectives are twofold. First, we want to gain insights into what drives consumers' consumption and purchase decisions. Second, we aim to identify factors important to incorporate in developing a structural model, and others that might be reasonably abstracted away.

3.1. Empirical Setting and Data Description

An anonymous online movie rental service in the United States (henceforth the "focal firm") provided the data on the condition of confidentiality. The focal firm operates on the same business model as Netflix. It uses U.S. Postal Service first-class mail to send its subscribers DVDs, along with postage-prepaid envelopes for the return of DVDs. The focal firm offers niche content of family-friendly movies and, with fewer than 100,000 subscribers, it is small compared with Netflix. It offers four regular subscription plans, with mailing quotas of one, two, three, and five, respectively. We focus on the three most popular plans with quotas one, two, and three. We refer to these plans as the "Low plan," "Medium plan," and "High plan," respectively. The same set of plans was offered during the entire observation period, and there were no changes in the monthly subscription prices: $11.95 for the Low plan, $19.95 for the Medium plan, and $29.95 for the High plan. The firm's subscription policy is that by default, the consumer's current subscription plan will be renewed for the next monthly billing cycle. Furthermore, the consumer may only change her plan at the beginning of, but not within, a cycle.

The data set contains detailed information about the payment and shipping records for a representative sample of 400 consumers observed between August 2003 and May 2005, for approximately two and a half years. Each consumer is identified by a firm-assigned ID and a (partial) credit card number. The payment history records for each consumer include the amount and date. By matching the payment sequence and the menu of plans offered, we reconstruct the purchase sequence (i.e., plan choices) of a consumer. We use the first payment date to determine when the consumer joined the service and the last payment to determine consumer exit.

The focal firm delivers to consumers nationwide from a *single* distribution center in a mountain state. Thus, there is natural variation in the service time across consumers, based on the distance between the distribution center and the consumer's mailing address. Specifically, the firm characterizes the service time as either five days for consumers who are geographically close or seven days for those who are located farther away. The shipping history for a representative consumer contains the rental records for all rented movies during her subscription. Specifically, for each movie rental, we observe the exact dates when it was sent out and received by the firm. The rental records show that in almost all instances ($\geq$ 99.5%), the firm promptly shipped the same number of movies to the consumer on the next business day after it received the returned movies. Given that the consumer needs to populate her own movie queue, there is the potential issue that if the consumer has not added a sufficient number of movies to the queue, the firm would not be able to ship anything after it received a product from the consumer, leading to the closed loop being broken.[2] In practice, more than 95% of the consumers in our sample maintain a sufficiently large number of movies in their queues.

We infer the date when the consumer received a movie on the basis of when the movie was shipped out by the firm and the one-way shipping time. Similarly, the *date of consumption is inferred using the date when the movie was returned*: this assumes that consumers will return the movie immediately after watching it, an assumption that has been made in the literature (Milkman et al. 2009).

We construct the entire subscription and rental history for all 400 consumers, spanning approximately 113,000 daily-level observations, detailed in Table 1. For any given day when the consumer is an active subscriber, we know the plan she subscribes to, the number of days until the next payment, the number of movies that she has available for immediate consumption (movie inventory), and the number of movies in the mailing process. From Table 1, observe first that movie inventory and movies in the mail add up to the quota, at individual and aggregate levels. Average movie inventory is 70.2% of the average purchased quota, indicating service time can constrain consumer's usage. Second, the average daily consumption is low (0.103), and higher during weekends. Third, the Medium plan is the most popular, chosen in 73% of all time periods, followed by the Low plan (14%) and the High plan (13%). Fourth, consumers are approximately

Table 1. Summary Statistics

	Mean	Standard deviation	Min	Max
No. of consumers	400			
Quota	1.98	0.49	1	3
Movie inventory	1.39	0.83	0	3
Movies in mail	0.59	0.76	0	3
Consumption	0.103	0.36	0	3
Low plan	0.14	—	—	—
Medium plan	0.73	—	—	—
High plan	0.13	—	—	—
Service time	6.01	1.001	5	7
Tenure in days	283.0	238.4	30	989

Notes. Observation period: August 2002 to May 2005. Number of observations: 113,014.

equally split between shorter and longer service times. Finally, there is substantial variation in the number of days consumers subscribe to the service.

3.2. Evidence from the Data

We present model-free evidence on consumption and purchase decisions over time. We first consider a number of likely seasonality (monthly and weekday/weekends) and heterogeneous viewing preferences. We then examine whether consumers demonstrate consumption smoothing or bunching.

3.2.1. Effects of Movie Inventory and Service Time on Consumption Rates. Because of the closed loop process, the mailing quota is the sum of (1) movie inventory and (2) movies in the mailing process, which will be received later by the consumer. Observe that the consumer's maximum daily-level consumption is restricted to her movie inventory, rather than quota. The consumption probability varies according to (1) quota, (2) the size of the movie inventory (zero, one, two, and three), and (3) the individual-specific service time (five or seven days).

A few observations follow from Table 2. First, the overall consumption rate is low. For example, consumers do not watch any movie more than 85% of the time even with multiple movies available. Second, for each of the service times, the probability of having positive consumption generally increases with the size of the movie inventory. Third, comparing consumers with different service times, we find that conditional on the number of movies available, the longer the consumer needs to wait (i.e., a longer service time), the lower the consumption probabilities. The same pattern is observed for all three levels of nonzero inventory sizes. These observations indicate that the consumer's usage is affected by both the inventory and service time.

3.2.2. Seasonality. To investigate whether seasonality impacts consumption, we examine it at the week and month levels of aggregation. Figure 1(a) suggests that at the aggregate level, the weekend consumption might be systematically different from the weekday consumption. Given our empirical setting, this impact might be expected and should be incorporated in the model.

To examine whether similar monthly variations exist, we compute the average consumption for each calendar month. We find that the monthly average consumption is quite stable, with a mean (standard deviation [SD]) consumption of 2.99 (0.07). Given the low coefficient of variation, 0.024, we do not find monthly seasonality to be an important driver for consumption rates.

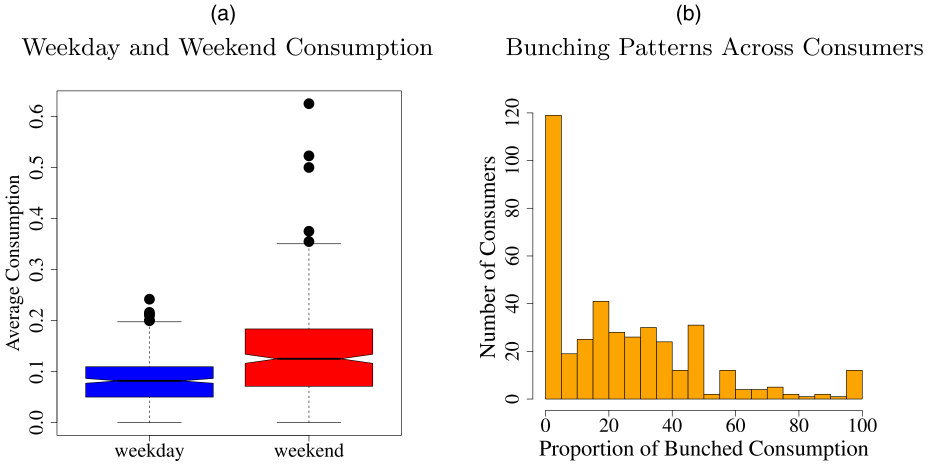
3.2.3. Heterogeneous Viewing Preferences. Consumers may be heterogeneous in their viewing preferences across movie genres, so they may hold on to certain movies for a longer time before watching them (e.g., Milkman et al. 2009). To investigate this effect, we examine movie title information in the shipping records for more than 11,000 movies sent out to the 400 consumers. On the basis of the movie titles, we categorize the movies into nine main genres (Action, Children, Classics, Comedy, Drama, Romance, Sci-fi, Suspense, War), which jointly accounted for more than 99% of all movies delivered. We also find that consumers watch an average of 5.9 genres. We then compute the number of days each movie was kept by the consumer as the time elapsed between when she received the movie and when she returned it, which is used as a proxy for utility. In the results in Table D.3 in Appendix D, we find no evidence that genre impacts how long consumers hold DVDs.

Table 2. Data: Inventory and Consumption

	Service time							
	5-day				7-day			
	$c_{it} = 0$	$c_{it} = 1$	$c_{it} = 2$	$c_{it} = 3$	$c_{it} = 0$	$c_{it} = 1$	$c_{it} = 2$	$c_{it} = 3$
Inventory = 0	100.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0
Inventory = 1	89.6	10.4	0.0	0.0	92.4	7.6	0.0	0.0
Inventory = 2	87.5	9.4	3.2	0.0	91.5	6.0	2.5	0.0
Inventory = 3	84.9	10.2	3.3	1.5	91.3	5.2	2.1	1.4

Note. Values are percentages.

Figure 1. Consumption Patterns from the Data



3.2.4. Consumption Bunching or Smoothing. A large movie inventory affords the consumer the opportunity to either engage in consumption bunching (i.e., watching multiple movies on the same day) or smoothing (i.e., watching one movie in a day). We investigate whether consumers are homogeneous in their preferences for consumption bunching. For each consumer, we compute the proportion of bunching (i.e., instances in which the consumer has either two or three movies available *and* chooses to watch at least two movies). Figure 1(b) shows the distribution of consumption bunching. Across all consumers, bunching accounts for 24.8% of all feasible occasions. Further investigation reveals substantial heterogeneity: the range is between 0 and 100%, and the standard deviation is 24.7%. A majority of consumers do not bunch their consumption. This result implies that it is important to account for bunching and smoothing as an important dimension of consumer heterogeneity.[3]

4. Model

In this section, we develop an integrated model of consumers' endogenous purchase and consumption. Consumers first make a *long-term* decision (i.e., choosing a plan or discontinuing the service at the beginning of each subscription cycle, which is a month for the focal firm). Next, consumers make a *short-term* consumption decision every period (day) on how many movies to watch.

We denote the set of payment or billing time periods using $\mathbf{T_p}$, and the set of weekend time periods (i.e., Saturdays and Sundays) as $\mathbf{T_w}$. Consumers in the model are indexed by $i \in \mathbf{I}$, plan choices by $q \in \mathbf{Q}$, consumption choices by $c \in \mathbf{C_t}$, and time periods by t. Plans in our setting correspond to a quantity indicating the maximum number of movies the consumer can check out at any time. The plan choice set $\mathbf{Q}$ is fixed across time, and consumers can make plan choices at regular payment periods, (i.e., $t \in \mathbf{T_P}$). Consumers can choose to change the plan or leave the service (outside option): $q \in \mathbf{Q} = \{0, 1, 2, 3\}$, where 0 represents the outside option. In a nonpayment period, $t \notin \mathbf{T_p}$, consumers cannot change their plan. The unit of analysis in the model thus incorporates both the plan and consumption choices available to a consumer at any time period. For notational convenience, we drop the i subscript on parameters.

4.1. Short-Run Period Utility of Consumption

The instantaneous period utility for consumer i in period t when choosing a decision c_{it}, denoted as $u(c_{it})$, is specified as a linear-quadratic form.[4] The linear-quadratic utility function is a popular choice model for consumers' usage decisions in other settings with subscription plans (e.g., Iyengar et al. 2007, Lambrecht et al. 2007).

$$u(c_{it}) = \alpha_1 c_{it} + \alpha_2 c_{it}^2 + \alpha_3 c_{it} \nu_{it} + \underbrace{\alpha_w c_{it} \ \mathbf{I}[t \in \mathbf{T_w}]}_{\text{Weekend Effect}} + \underbrace{\alpha_{cs} c_{it} \ \log(\omega_t)}_{\text{Content Set}}. \tag{1}$$

The first two terms with coefficients α_1 and α_2 denote the linear and quadratic terms of the utility function.[5] The third term incorporates the stochastic term or shock, ν_{it}, interacted with the consumption utility. The single-dimensional shock can be thought of as impacted by amount of leisure time, or movie viewing preferences. We specify the distribution of this shock through a positive, continuous distribution, in practice the log normal. Higher values of the shock would then imply higher marginal utility from consumption, if we have $\alpha_3 > 0$. Consumers who prefer to intertemporally smooth their consumption would thus have relatively high values of α_1 and low values of α_3. In contrast, higher values of α_3 would lead to higher preference for bunching.

The fourth term indicates the weekend effect, whereby consumers can receive a different (possibly higher when $\alpha_w > 0$) utility for watching movies during the weekend, and is also multiplicative in the consumption. The last term is intended to capture how a consumer's utility could vary with the content set, the size of the content library available from the firm. Note that content set utility is interacted with the consumption amount c_{it}, so it only accrues when a consumer watches at least one movie in that period (i.e., $c_{it} \neq 0$). Higher positive values of α_{cs} would indicate consumers having higher marginal consumption utility when the content set is large, which can occur when a consumer is more likely to find a better match with her preferences from a larger content set. We model this through a concave function (log) in practice, which allows for increasing utility from more content with a diminishing marginal impact.

We have examined the static instantaneous (period) utility of a consumer in the above discussion. If consumers were myopic (i.e., static utility maximizers), then they would make consumption choices to maximize the function above. However, this does not account for the fact that the current decisions of consumers have a significant impact on their future utility. Because the future sequence of consumption opportunities (choice set) depends on the current period decision, it is important to understand these trade-offs.

4.2. Short-Run Dynamics

The basic intertemporal trade-off in consumption is the following: suppose a consumer watches all movies in her possession during the current period. Then, she has to wait for τ days (service time) to receive new movie titles. This would imply that if she received high shocks, ν_{it} during the intervening periods before she receives new titles, she has to forgo those consumption opportunities. To formalize this, we define the short-term state space. The state for consumer i in period t has four components and is defined as s_{it}: $s_{it} = (x_{it}, w_{it}, z_{it}, \omega_t)$. The components of the state variable are as follows: x_{it} indicates the mailing state of the consumer, w_{it} denotes the day of the week, z_{it} tracks the time (number of days) to the next cycle date, when the consumer is allowed to change plans, and, finally, ω_t captures the content set available. The first three components of the state space have deterministic transitions, whereas the last state variable denotes the content set and is modeled as constant. We examine the evolution of each component in turn next.

4.2.1. Mailing State. Consumers watch received movies, return them, and obtain new movies after the service time. Formally, we define the service time τ as the time it takes for a consumer who mails a movie in period t to receive a new movie, that is, in period $(t + \tau)$. The service time includes both the two-way mailing time and processing time. It is therefore exogenous to the consumer and partly determined by the delivery service (e.g., the postal or courier service).

We introduce the mailing state x_{it}, which fully characterizes the closed loop rental process and the consumer's dynamically evolving consumption set. This vector details the state of each product (movie) through the process of obtaining a movie, watching it, and returning it to the firm, which in turn processes the returned movie and mails out the next movie to the consumer. The service time critically determines the state space. The mailing state for consumer i in period t is

$$x_{it} = \left(x_{it}^0, x_{it}^1 \ldots, \underbrace{x_{it}^s}_{\substack{\text{Number of movies} \\ \text{expected in period } (t+s)}}, \ldots, x_{it}^{\tau_i} \right), \qquad (2)$$

where τ_i is the service time for consumer i. Note that x_{it}^0 is the number of movies currently held by the consumer, whereas x_{it}^s denotes the number of movies that the consumer will receive in the future period $(t + s)$, where $s \in \{1, 2, \ldots, \tau_i\}$. Observe that the total number of movies across all of the states $(x_{it}^0, x_{it}^1, \ldots, x_{it}^{\tau_i})$ must be equal to the number of movies in the plan: $\sum_{s=0}^{\tau_i} x_{it}^s = q_{it}$.

Although the service time is exogenous, the transition process for x_{it} is *endogenous* and determined by the consumption and plan choices made by the consumer. We detail the mailing state transition in Appendix C.1. Importantly, the consumption decisions are intertemporally linked owing to the consumer's current inventory level, which imposes two constraints for the consumer. The first and explicit constraint is that the inventory imposes a hard cap on current consumption. The second and less explicit constraint is through the trade-off consumers must make between current and future consumption opportunities. Given the uncertainty, the consumer has to determine a consumption plan to optimize the total utilities from both immediate consumption and consumption in the near future. To capture consumers' intertemporal trade-offs, we need to understand how current choices made by the consumer will impact future choice sets. We conceptualize two different sources of intertemporal trade-offs, expanded in detail below.

The first intertemporal trade-off evaluated by the consumer is between watching the available movie(s) now versus later. The intuition becomes apparent as we consider a consumer with the smallest plan, $q = 1$ and a service time of $\tau_i = 5$ days. When the consumption utility is low owing to an idiosyncratic shock, the consumer is likely to derive higher utility from postponing consumption. In other words, waiting provides an *option value* for the consumer. Such an intertemporal trade-off is driven by the *uncertainty* in consumption utility, which is not possible for the consumer to perfectly predict owing to uncertain factors, such as available time for consumption.

Specifically, at time period t, the consumer receives an idiosyncratic consumption shock of ν_{it}. If the shock is sufficiently high (e.g., due to unanticipated time availability or a suitable occasion), then the consumer will have the following trade-off. She could watch the movie in period t but would have to wait for $\tau_i = 5$ days for the next movie to arrive. Thus, even if she has a better occasion (i.e., a higher idiosyncratic shock) to watch movies during the days $\{t+1, \ldots, t+5\}$, she will not have movies to watch. Thus, the consumer would want to wait for a sufficiently high level of shock to choose in order to consume within the current period.

Second, in addition to the above trade-off, consumers with plans $q > 1$ account for the number of movies that are expected to arrive in the near future when making their consumption decisions. Consider a consumer who subscribes to a three-movie plan ($Q = 3$) and two situations: (a) she either has a high current inventory (three movies), with no movies in the mail or (b) a low current inventory (one movie), with two movies scheduled to arrive in the mail after five days. Will she be equally likely to watch *one* movie in each of the situations (a) and (b)? We note first that she derives the same amount of immediate utility from immediate consumption under both (a) and (b). However, there is a difference due to the intertemporal trade-off. If she watches a movie in scenario (a), she will still have two movies in her inventory if a sufficiently high consumption occasion arises the next day. On the other hand, if she watches a movie in scenario (b), she will not have any movies to watch for the next five days until a movie arrives. Thus, she will likely miss high-value consumption opportunities that occur in the near future.

More broadly, consumption decisions will be different because they lead to different future options. Because consumers value the (discounted) future utility, as well as the immediate utility, it is rational to reduce consumption when the current inventory is low and increase it when inventory is high. The above mechanism extends to consumers across multiple plans, but the level of shocks required to make a current consumption decision will be different across different states, as well as across different plans for the same consumer. Furthermore, the above arguments suggest that a shorter (longer) service time reduces the value of waiting for a forward-looking consumer, and consequently increases (decreases) the likelihood for immediate consumption.

4.2.2. Additional Short-Run States. We include a *weekend state* variable that captures the day of the week. It is the day of the week variable $w_{it} \in \{1, 2, \ldots, 7\}$, beginning with Monday (day 1), and ending with Sunday (day 7). We detail the weekend state transition in Appendix C.2.

The next state variable is the *time from the billing period* z_{it}, which takes values from the set $\mathbf{Z}$, $z_{it} \in \mathbf{Z} = \{1, \ldots, T\}$, where T is the length of the billing cycle. Recall that $\mathbf{T_p}$ denotes the set of all payment or billing periods. In periods with $z_{it} \neq T$, consumers make only consumption choices, whereas in periods in which $z_{it} = T$, they make choices on both consumption and the plan.[6] In the short run the content state ω_t is fixed, and we examine its evolution in the long-run dynamics discussion below. z_{it} is reset to 1 on payment period and increments by 1 (i.e., $z_{it} = z_{i,t-1} + 1$ on nonpayment periods).

A forward-looking consumer solves the following short-run T-period consumption problem:

$$\max_{c_{i\tau} \in \mathbf{C}_{i\tau} \forall \tau \geq t} u_{it}(c_{it}, \nu_{it}) + \mathbb{E}\left[\sum_{\tau=t+1}^{t+T-1} \beta^{\tau-t} u(c_{i\tau}(\nu_{i\tau}), \nu_{i\tau})\right]. \qquad (3)$$

This finite-horizon dynamic program can be represented in terms of the Bellman equation of the period-specific ex ante value function V_t as a function of the state variable s_{it} defined earlier.

$$V_t(s_{it}) = \mathbb{E}_\nu\Bigg[\max_{c \in \mathbf{C_{it}}}\Big(u_{it}(c, s_{it}, \nu) + \beta\, \mathbb{E}_{s_{it+1}|s_{it}}\big[V_{t+1}(s_{i,t+1})|s_{it}, c\big]\Big)\Bigg]. \qquad (4)$$

Because this is the short run, we set the value function beyond the terminal period to be zero (i.e., $V_t = 0$ for $t \geq T$). We also use the expected value function at the beginning of the short run denoted as V_0 in the long-run dynamics.

Section 4 embeds two key trade-offs corresponding to short-run dynamics. The first is the trade-off between current and future consumption. A myopic consumer might choose to watch all movies when the immediate consumption utility for that option is the highest of all options, whereas forward-looking consumers might wait because they internalize the negative impact of watching all movies immediately, in that they would have nothing in the inventory until the service time has elapsed. The second is the trade-off between intertemporal smoothing and bunching. In particular, consumers with high α_3 and low α_1 have stronger incentive to bunch consumption, whereas consumers with high α_1 and low α_3 derive value from regular, periodic consumption. In Appendix B we show how the thresholds and choice probabilities are derived.

4.3. Long-Run Plan Choice

In the long run, consumers make trade-offs between high- and low-quota plans: high-quota plans charge higher prices, yet they not only allow consumers higher utility with more options for immediate consumption, but also diminish the likelihood that consumers will stock out during time periods with a high idiosyncratic shock. Both immediate utility and greater flexibility in consumption become even more important as the content set increases, and consumers obtain higher consumption utility owing to the larger variety in product choices.

The period long-run utility function during payment periods $t \in \mathbf{T_p}$ follows:

$$U(q_{it}) = V_0(s_{it}(q_{it}, \omega_t)) + \alpha_p p(q_{it})\epsilon_{it} + \alpha_{sw} I[q_{it} \neq q_{i,t-1}]. \quad (5)$$

The above period utility incorporates the short-run value function V_0 defined above as the short-run value function at the beginning of a plan period (month). Both the short-run value V_0 and price depend on the plan choice q_{it} made by the consumer. The price coefficient is α_p (expected to be negative), and the shock ϵ_{it} impacts the disutility of price multiplicatively. Thus, we use a single scalar shock to rationalize the plan chosen by the consumer. If $\alpha_p < 0$, when ϵ_{it} is high the consumer is more likely to choose a plan with a lower price, everything else equal.

An alternative would have been to just use a multinomial choice model. If we had used a multinomial logit specification, say, with separate shocks for each plan, then the model would ascribe positive probability to plans that are dominated. We specify the shock ϵ_{it} using a log normal distribution independent and identically distributed across consumers and time periods. The final term represents the switching costs. We model the idea that consumers face a switching cost whenever they choose a plan that is different from the previous choice (i.e., $q_{it} \neq q_{i,t-1}$). This term captures the common feature of subscription services (and also our setting), such that if the consumer does not make an active plan choice change in the billing period, then she retains the previously chosen plan. Similar to other research, we do not attempt to ascribe microfoundations to the switching costs (Goettler and Clay 1997, Shum 2004, Handel 2013). As an example, switching costs could be interpreted as the time and effort cost of logging into the firm's website to switch to a different plan, though the firm does not explicitly charge a fee to change plans. Note that the consumer may discontinue the service by choosing plan 0 (i.e., $q_{it} = 0$) and will incur the switching cost as well.

4.4. Long-Run Dynamics

In the long run, the above utility specification includes intertemporal trade-offs in plan choice for the consumer involving the content set, prices, and switching costs. Recall that in choosing a plan, consumers account for the expected sum of discounted utilities for consumption choices that are enabled by a specific plan. Thus, consumers who have a higher consumption utility will, everything else being equal, tend to choose higher plans in the model. Additionally, as the size of the content set changes, consumers will be more inclined to upgrade to higher plans, because they might obtain higher utility from consumption. Finally, consumers face a switching cost in changing plans, which can cause them to continue with their current plan, even though a higher plan might become more attractive because of a larger content set. To capture these trade-offs, we define the long-run state $S_{it} = (w_{it}, q_{it}, \omega_t)$ to include the day of week, the current plan choice, and the content state. The first two are deterministic, whereas content state is stochastic as described below.

4.4.1. Content Set State. The final state variable ω_t denotes the size of the content set or the number of movies available in period t. From the consumer's perspective, it evolves exogenously as a random process. Note that ω_t is the same for all consumers. Consumers form expectations about the stochastic evolution of the content set, and we specify this state variable by a probability distribution across a discrete number of content set sizes, $\omega \in \{1, 2, \ldots, N_\omega\}$, and with a $(N_\omega \times N_\omega)$ probability transition matrix $\mathbf{\Omega}$. Consumers have rational expectations and will expect the content set size to evolve according to $\mathbf{\Omega}$.[7] We let the content set directly impact the consumption utility in Equation (1), so that consumers obtain an increasing utility with a larger content set, but with diminishing marginal utility.[8]

The infinite-horizon dynamic program can be represented in terms of the Bellman equation of the ex ante value function W as a function of the state variable S_{it}:

$$W(S_{it}) = \mathbb{E}_\epsilon\Big[\max_{q \in \mathbf{Q}} \Big(U(q, S_{i,t}, \epsilon) + \beta^T\, \mathbb{E}_{S_{it+1}|S_{it}}\big[W(S_{i,t+T})|S_{it}, q\big]\Big)\Big]. \quad (6)$$

Equation (6) embeds the trade-off in long-run dynamics. The time-frame for long-term decisions is every T periods, so the future utility is discounted by β^T.

5. Identification and Estimation

We detail how consumer preference parameters are separately identified and the variation in the data that aids in such identification. Then, we provide an overview of the estimation process (a detailed algorithm is provided in Appendix C).

5.1. Identification

We begin with the consumption parameters that can be identified with daily consumption patterns. First is the discount factor. Note that the identification of the discount factor is well known to be generically confounded in dynamic discrete choice models without an exclusion restriction (Magnac and Thesmar 2002). We do not attempt to estimate it, and rather set the daily discount factor at $\beta = 0.999$ for all consumers, as in most of the literature in dynamic discrete choice models.

Next, we examine the consumption parameters α_1, α_2, and α_3, which we identify only using short-run consumption data. We have to normalize one of these parameters for identification, and given our empirical interest in understanding heterogeneous consumers who have either different baseline consumption utility or high variance in consumption utility, we normalize $\alpha_2 = -1$, while estimating α_1 and α_3.

The consumption parameter estimates are determined by the average consumption rates as well the variance of consumption in terms of bunching versus smoothing. The average consumption frequency during periods when the consumer has at least one movie available identifies α_1. Among consumers with the same average rate of consumption, some might show higher variance in day to day consumption, which would lead to higher values of α_3.

For the weekend coefficient, α_w, identification is provided by the difference in the consumption probability during the weekends, compared with weekdays, with a larger value of the coefficient indicating a higher difference in consumption probabilities.

For the long-run parameters, we specify the price and switching cost as homogeneous. The price coefficient, α_p, is identified by the consumer choice of the plan, relative to their long-term average consumption. Intuitively, the consumption parameters determine the expected utility consumers obtain from one month of consumption, and the price coefficient will determine how consumers on average determine the price equivalent of these utilities and make plan choices. Our plan choice and prices do not vary over time, and if the data had such time series variation, that would permit us to obtain heterogeneous price coefficients.

Now consider the switching cost coefficient. Conceptually, the consumer's willingness to pay for consumption can be identified according to consumption and plan choices during the periods before the first change in the content set. Then, as the content set increases, the switching cost can be identified by the consumers' observed plan-switching decision. When the content set increases, if the consumer has higher value for a plan with a higher quota compared with their current plan, but chooses to stay with their current plan instead of switching to the higher plan, we attribute that to switching costs. The willingness to pay for all plans can be determined owing to the known form of the consumption utility function. We note that because the consumption utility is not modeled as genre-specific, a potential confounding factor for the switching cost is the consumers' heterogeneous viewing preferences across movie genres. We discuss this issue in Appendix D.

The firm's pricing structure has not changed significantly during our observation period. Note that although the prices are set according to the quantity of each plan, we assume that each plan does not have separate fixed effects. Specifically, consumer value for each plan is only due to the consumption opportunities enabled by the plan, and there is no separate value in purchasing a plan. Because we can obtain consumption preference parameters only on the basis of consumption data, we can obtain the normalized value or willingness to pay a consumer has for each plan.

5.2. Estimation

The discrete-choice, dynamic structural model developed above captures both the static and intertemporal trade-offs faced by consumers. Estimation of this model, however, is computationally challenging for three reasons. First, the value function is highly jagged owing to the multidimensional mailing state. Second, the dimension of the payoff-relevant state variables is large.[9] Third, the value function iteration takes longer to converge when the discount factor is closer to one, as in our case: because our time period is a day, this leads to a high discount factor ($\beta = 0.999$), and the problem is significantly compounded.

We use a hierarchical Bayes (HB) estimation approach to allow for individual-level heterogeneity in consumption. HB models are known to reliably capture aspects of a wide variety of data generating process (Andrews et al. 2002). Although HB methods have been commonly used in marketing (Allenby and Rossi 1998), they are relatively rarely used for models with forward-looking consumers. The primary reason is computational, because in dynamic structural models, we need to typically solve the Bellman equation and obtain the value function by performing value function iteration to convergence for each value taken by the parameters in the estimation process. Bayesian methods typically require thousands of iterations across the parameter space to achieve convergence, making it challenging to combine them. Imai et al. (2009) (IJC) propose a novel and highly practical method to reduce the computational burden by interweaving the Markov chain Monte Carlo (MCMC) iteration with only one step of the value function iteration. They demonstrated that such an iterative process will converge to the correct posterior distributions and provide convergence of the value function as well. The crucial aspect of their method is that value function iteration until convergence is *not* required at each MCMC iteration, and the algorithm efficiently uses past information to form approximations of the true value function.

Consumers are heterogeneous in their valuation of consumption quantity, with parameters (α_i^1, α_i^3). These parameters have the specification: $(\alpha_i^1, \alpha_i^3) \sim \mathbf{N}(\Delta, \mathbf{V}_\alpha)$.

The priors are specified as $\mathrm{vec}(\Delta|\mathbf{V}_\alpha) \sim N(\bar{\delta}, A_\alpha)$, where $A_\alpha = A^{-1}V_\alpha$ and the prior on the covariance matrix is $V_\alpha \sim IW(\nu, I_{|\alpha|})$. We specify uninformative priors in our empirical implementation: although the Bayesian framework allows the researcher flexibility to incorporate prior information, we do not have additional information regarding consumer demographics or preferences beyond their consumption and purchase histories. We provide the detailed estimation procedure in Appendix C.

Although the IJC algorithm uses smoothing utilizing a kernel regression, this smoothing is across the parameter space rather than the state space. Therefore, the jagged nature of the value function across the state space will not result in a problematic approximation for IJC, when compared with methods that interpolate the value function over the state space. Although the IJC algorithm substantially alleviates the computational burden and allows us to account for individual-level heterogeneity, the large state space in our research setting nevertheless renders the estimation time-consuming.[10] We use a sample of $N = 200$ consumers in estimation.[11]

6. Results

We estimate four alternative models for comparison. The first, model 1, assumes homogeneous consumers. Models 2, 3, and 4 incorporate consumer heterogeneity but differ in other ways. Model 2 uses a semiparametric utility function, whereas model 3 uses a linear-quadratic formulation for the parametric forms (linear or quadratic) of consumption utility. Finally, model 4 leaves out switching costs (i.e., setting it to zero).

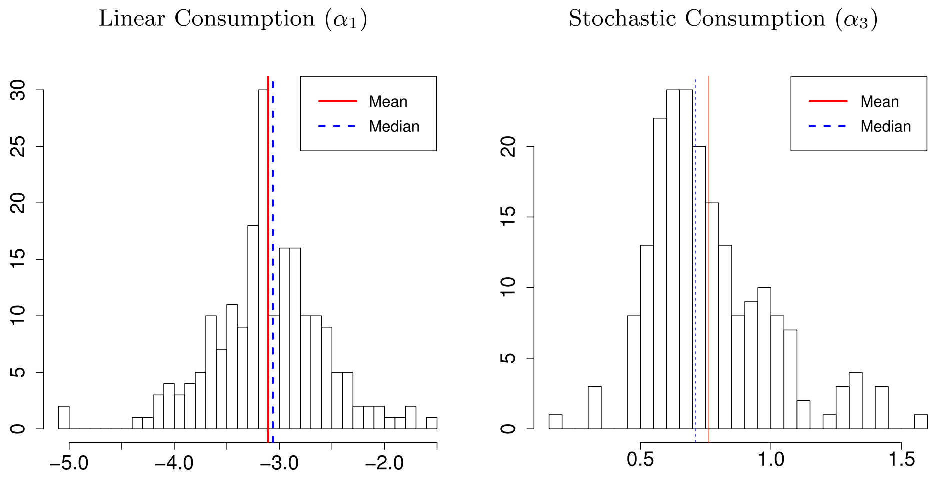
Following the convention in the literature (e.g., Rossi et al. 2012), we use log-marginal density (LMD) as the basis for model comparison. We also use the Bayesian information criterion (BIC) for model comparison with inclusion of a notion of model complexity. In the last rows of Table 3, we report the LMDs and BICs of the four alternative models. The linear-quadratic heterogeneous model (model 3) shows significantly lower values across the models, based on LMD and BIC. We summarize the model estimates in Table 3, focusing on the means and highest posterior density of the posterior distributions of the parameters. For the heterogeneous parameters, the hierarchical means are reported.

Table 3. Parameter Estimates

	Model (M)			
	M1, Homogeneous	M2, No switching cost	M3, Linear-quadratic	M4, Semiparametric
Linear consumption $(\alpha_{1,i})$	−1.045 [−1.143, −0.964]	−1.505 [−2.083, −1.231]	−3.106 [−3.507, −2.760]	
Quadratic consumption $(\alpha_{2,i})$	−1	−1	−1	
Stochastic consumption $(\alpha_{3,i})$	0.686 [0.659, 0.713]	0.423 [0.388, 0.462]	0.762 [0.696, 0.830]	+1
Consumption $c = 1$ (θ_1)				−0.943 [−0.984, −0.902]
Consumption $c = 2$ (θ_2)				−0.250 [−0.284, −0.217]
Consumption $c = 3$ (θ_3)				0.646 [0.619, 0.680]
Weekend (α_w)	0.028 [0.021, 0.035]	0.026 [0.019, 0.032]	0.041 [0.026, 0.063]	0.016 [0.015, 0.018]
Content set (α_{cs})	0.046 [0.037, 0.057]	0.182 [0.140, 0.270]	0.341 [0.283, 0.403]	0.869 [0.669, 1.121]
Price (α_p)	−0.154 [−0.170, −0.141]	−0.242 [−0.255, −0.229]	−0.201 [−0.222, −0.180]	−0.368 [−0.402, −0.344]
Switching cost (α_{sw})	−1.875 [−2.077, −1.736]		−5.822 [−6.443, −5.100]	−3.685 [−4.023, −3.445]
-Log marginal density	18,974.5	19,602.2	18,519.2	20,021.9
AIC	37,967.0	39,220.4	37,056.4	40,063.8
BIC	37,959.9	39,215.3	37,049.3	40,054.7

Notes. For heterogeneous parameters, posterior mean of the hierarchical (population) parameter are reported. 95% highest posterior density (HPD or credible) intervals are reported in brackets below the estimates.

Figure 2. Individual Parameters from Baseline Model


Linear Consumption (α_1)

Stochastic Consumption (α_3)

Examining the consumption parameters in the baseline linear-quadratic model (model 3), we find that the posterior mean consumption utility parameters $\alpha_1(<0)$ and $\alpha_3(>0)$ have the signs we might expect from the model-free choice probability values, because consumption has a lower probability than nonconsumption.

We model consumption at the daily level, and negative parameters reflect the data that in most periods (days), consumers chose not to watch movies, even when they have movies in their inventory. This pattern is due to the higher utility they place on the outside option, which incorporates alternative, non-movie-watching activities. Thus, the consumer would only choose to consume in periods when they receive a sufficiently positive utility shock. Recall $\alpha_2 = -1$ for identification.

The weekend effect is positive, suggesting consumers on average prefer to consume on weekends, compared with weekdays. Combining the parameter estimates for the consumption coefficients discussed above, we find that although the positive weekend consumption effect partially offsets the average negative consumption utility, it does not affect the ordering across the choices in the main model specifications. The content size parameter is positive, suggesting that consumers derive additional value from a larger content set, as one might expect, owing to a better match of content to consumers' preferences. None of the credible intervals for the short-run parameters includes zero.

Focusing on the long-run parameters, the population price coefficient is estimated to be negative across all specifications. The switching cost is also estimated to be negative, reflecting the cost consumers face to change plans. To better understand the magnitude of switching costs, we combine estimates for price sensitivities and switching costs to compute the monetary equivalence of the switching cost by normalizing it with respect to the price coefficient $\left(\frac{\alpha_{sw}}{\alpha_p \exp(\frac{1}{2})}\right)$, where $\exp(\frac{1}{2})$ is the mean of the standard log-normal distribution. We find the switching cost to be \$17.57. The magnitude of switching cost is somewhat in line but typically lower than found in prior research (e.g., Goettler and Clay 1997).

Figure 2 illustrates the distributions of the *individual-level* posterior means for each of the heterogeneous parameters of the main linear-quadratic specification (model 3), with both the population mean (solid red line) and the median (dotted blue line) shown. We observe that overall, there is significant individual-level heterogeneity.

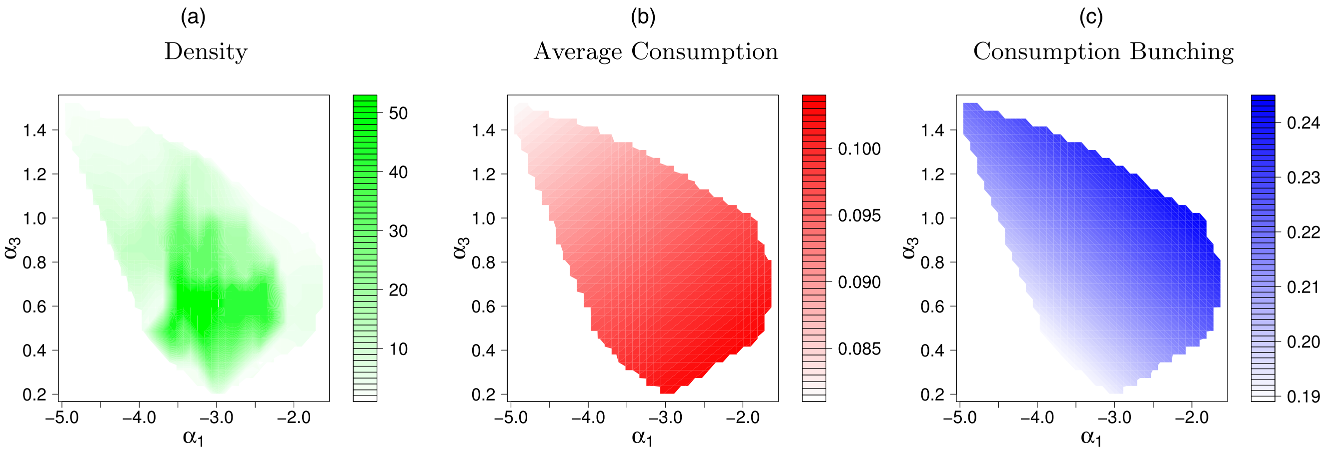
Table 4 details the hierarchical parameter V_α, which captures the variance and correlation of the heterogeneous parameters. We find a negative correlation between α_1 and α_3. A primary issue was to understand bunching versus smoothing consumption patterns, and Figure 3 connects the estimates of our structural model with the data patterns. In panel (a), the joint density plot of heterogeneous parameters, note the significant heterogeneity in both parameters around the posterior mean. This also conforms to the bunching and smoothing patterns in Figure 1(b).

Panel (b) shows how the average consumption varies according to the consumption parameters $\alpha_{1,i}$ and $\alpha_{3,i}$. Two observations follow. First, fixing $\alpha_{3,i}$ (i.e., moving horizontally from left to right), consumers with higher $\alpha_{1,i}$ have high consumption, conforming with intuition. However, fixing $\alpha_{1,i}$ and moving vertically from bottom to top, consumers with higher $\alpha_{3,i}$ have *lower* consumption frequencies. We might have expected the opposite, because consumption utility is increasing

Table 4. Posterior Mean and HPD of Hierarchical Parameter V_α (Baseline Model)

	α_1	α_3
α_1	0.609 [0.425, 0.854]	−0.123 [−0.192, −0.068]
α_3	−0.123 [−0.192, −0.068]	0.112 [0.084, 0.146]

Figure 3. Heterogeneity and Consumption Characteristics


(a) Density

(b) Average Consumption

(c) Consumption Bunching

in both parameters. However, recall that a high value of $\alpha_{3,i}$ leads consumers to wait for a high consumption shock (ν_{it}) term to be drawn to consume, and consequently, they consume less on average. More broadly, we note that it might be difficult to capture such heterogeneity by alternative ways of modeling consumer heterogeneity.

6.1. Elasticity

Because subscription services routinely offer a menu of plans (see Table A.1 in Appendix A for additional examples), it is useful to examine the purchase elasticities at the current price levels, using parameter estimates from the dynamic linear-quadratic model.

Table 5 details the elasticities (computed as arc-elasticities), where the first row indicates the change in purchases with respect to a price change in the Low plan, etc. First, we find that all own-elasticities are negative, confirming our expectation that consumers reduce purchases in response to a small price increase for the plan. Second, own-price elasticity is ordered from most negative to least negative as we move down the diagonal (i.e., from the Low to the High plan). This pattern is consistent with sorting (i.e., consumers who choose the High plan are less price sensitive).

Third, we find that all cross-elasticities are positive, indicating that consumers substitute across plans. The magnitude of these cross-elasticities suggests that a price change of the Low plan has a significant impact on the Medium plan, but the effect is negligible for the High plan. Note that the baseline purchase quantities are much higher for the Medium plan than for the Low or High plan; a proportional change in the quantity of the Medium plan will be larger in actual quantity change, compared with the Low or High plan.

Table 5. Price Elasticities of Plans

	Low	Medium	High	Outside option
Low	−5.81	0.32	0.016	0.67
Medium	0.52	−1.30	0.52	0.58
High	0.018	0.86	−0.78	0.01

Finally, for a price change in the High plan, we find that it has a negligible impact on the Low plan but a large impact on the Medium plan (i.e., consumers are likely to switch to the Medium plan as a substitute for the High plan). Consistent with our expectations, the outside option probabilities are most impacted by changes in the Low plan, followed by Medium and High plans.

7. Counterfactuals

The estimated structural parameters allow us to investigate the impact of service time on value creation and capture across a variety of pricing strategies. We show how consumer decisions on purchase and consumption are significantly impacted by service time. We examine a number of counterfactuals in this section; in all counterfactuals, we focus on the highest content state and set switching costs to be zero. Marginal costs for all counterfactuals are set at \$2, except when we explicitly consider the case of higher marginal costs.[12]

7.1. Service Time Reduction at Current Prices

As an illustrative example, we let the focal firm maintain its current prices and product portfolio. We reduce service time to five days for all consumers, so consumers located further away also have the same shorter service times. We find revenue increases by 6.4%, and costs by 13.7%, which results in profits increasing by 2.4%. To understand how heterogeneous consumers contribute, we create three-dimensional plots, where each point in the x-y plane represents the posterior mean $(\alpha_{1,i}, \alpha_{3,i})$ of consumer i, and the z-axis represents the change in revenue, costs, or profits. We define $\Delta Revenue([5,7] \rightarrow [5,5]) = Revenue(\tau = [5,5]) - Revenue(\tau = [5,7])$, so a positive value indicates a consumer increasing spending, with cost and profit terms defined similarly. The regression hyperplanes illustrate how each quantity varies with respect to $\alpha_{1,i}$ and $\alpha_{3,i}$.

Panel (a) of Figure 4 shows that although the revenue contribution from a majority of consumers increases, the revenue for a significant minority (23.5%) of consumers drops. The reductions in revenue are mainly attributed to consumers with a high $\alpha_{1,i}$ but a low $\alpha_{3,i}$. Consumers with high $\alpha_{1,i}$ value regular, frequent consumption and have an incentive to purchase more expensive plans to counteract the long service time. With a reduced service time, they are able to consume regularly using a lower plan, causing them to downgrade. In contrast, consumers with high stochastic utility ($\alpha_{3,i}$) avoid downgrading because they lose the ability to bunch consumption. In fact, some upgrade because the reduced service time allows them more bunching opportunities.

Panel (b) of Figure 4 shows significant heterogeneity in costs, which increase for 76% of all consumers. Consumers with a high $\alpha_{1,i}$ and low $\alpha_{3,i}$ downgrade plans, limiting consumption and costs, but costs increase substantially for consumers with a high $\alpha_{3,i}$ who increase consumption. Panel (c) of Figure 4 shows that the average profit increases for 64.5% of all consumers. High-$\alpha_{1,i}$ and low-$\alpha_{3,i}$ consumers lead to a lower profit, whereas for high stochastic utility consumers, in both cases, revenue impacts dominate the cost impact.

7.2. Service Time Reduction with Price Reoptimization (Low Marginal Cost)

We now evaluate the impact when the firm is able to reoptimize prices, accounting for how much consumers value improvements in service time. The impact on consumer surplus a priori is not obvious, because faster service leads to more consumption opportunities and higher valuation and prices, but the firm can potentially capture a higher share of the created surplus with prices, leading to uncertainty in the impact of service time on consumer surplus.

We consider the case in which the firm reduces the service time for both types of consumers ($\tau = 5, 7$) by the same number of days. We note that by reducing the service time, the firm might induce a higher consumption level. To enable this higher consumption, we assume that the firm faces no supply constraints and faces a low marginal cost of $MC = \$2$, an amount suggested by the firm as covering two-way shipping and some handling charges.[13]

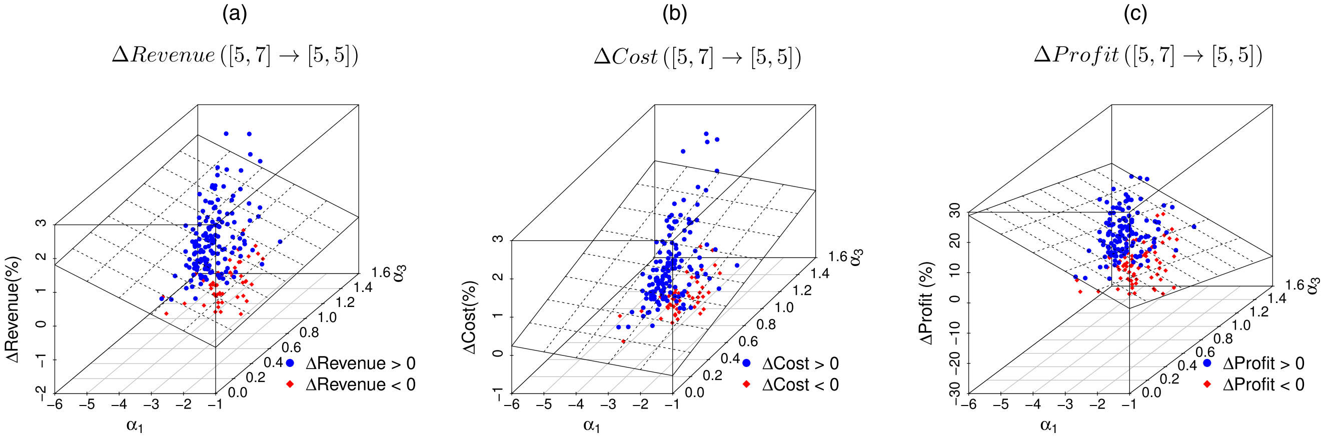
We assume that the firm maximizes overall profits across the price levels for the three plans, (p_1, p_2, p_3). The quantities corresponding to each plan (quota) are $Q = (Q_1, Q_2, Q_3) = (1, 2, 3)$ respectively. The objective function is

$$\max_{\boldsymbol{P}} \sum_{i=1}^{I} \sum_{m=1}^{M} \sum_{k=1}^{3} \beta^{mT} \times Prob(q_{ik}|\mathrm{P}) \times \left[\left(p_k - mc \times \sum_{t=1}^{T} E(c_{it}|q_{it}, \tau_i; \alpha_i) \right) \right], \tag{7}$$

where the firm chooses subscription prices $\boldsymbol{P} \equiv (p(q_1), p(q_2), p(q_3))$. Consumers are indexed by i, m denotes payment period (month), M the length of the simulation horizon (36 months), and T the number of days in a month (30).

Inside the bracket is the per-period profit of month m for plan $k \in \{1, 2, 3\}$, computed as revenue $p(q_{ik})$ net the total expected costs of serving consumer i. The marginal cost, denoted as mc, is assumed to be \$2 for each shipped movie. Note that the expected consumption amounts of consumer i depend on preference parameters and the mailing state (omitted for brevity), and the service time as well. Weighting the conditional profits with the corresponding probabilities of purchasing each plan, $Prob(q_{ik}|\mathrm{P})$, gives the expected profit contribution by consumer i in month m. The length of the time horizon, denoted as M, is set at 36 and corresponds to a three-year period.

Figure 4. Changes in Revenue, Cost, and Profit at the Individual-Consumer Level



(a)

$\Delta Revenue\,([5,7] \rightarrow [5,5])$

(b)

$\Delta Cost\,([5,7] \rightarrow [5,5])$

(c)

$\Delta Profit\,([5,7] \rightarrow [5,5])$

For each service time, we first compute the firm's revenue, costs, and profit per customer month and the average consumer surplus at the current price and service times as the basis of comparison (scenario 1: baseline). We then evaluate four alternative scenarios in which the firm retains the current service time (scenario 2, no improvement) or reduces service times for all of its consumers by one to four days (scenarios 3–6). For example, under a "small" improvement, consumers who had a five-day (seven-day) service time can now receive their new movies four (six) days after they return the old ones. For each scenario in 2–5, we let the focal firm optimize the prices for each of the three subscription plans. For ease of interpretation, we compute the percentage change with respect to the baseline scenario. The results are summarized in Table 6 below.

Comparing scenarios 1 and 2 (first and second columns of Table 6), we find that the optimized prices at the current service time are higher and lead to lower revenue (by 3.9%) but significantly lower costs (by 32.2%). Thus, the firm serves fewer consumers at higher prices, leading to a substantial reduction in consumer surplus (by 38.4%). Further comparisons among scenarios 3–6 show several interesting patterns. First, as the firm further reduces the service time, it enjoys greater pricing power, evidenced by the monotonic increases in the optimal revenue. Second, the firm incurs greater costs due to higher consumption levels. However, the profit increase is nonmonotonic. We first see profits increase as we reduce service time by up to three days, but it actually decreases with a further reduction in service time. This happens because the increase in consumer valuations is lower than the cost increase driven by consumption. The difference between valuations and costs diminishes, so the firm has lower ability to extract surplus. Consumer surplus, however, increases significantly, as we might expect.

7.3. Service Time Reduction with Price Reoptimization (High Marginal Cost)

Below, we consider consumer decisions and the firm's strategic choices when firm marginal costs are relatively high (set at \$4.5) from the baseline. This scenario represents additional costs incurred owing to inventory holding and replenishment, as well as fixed costs that can be made variable by business model choices made by the firms (e.g., paying for priority shipping).

In sum, with high marginal cost, an improvement in service time may create a *double whammy* for profitability: not only do the total costs of serving customers increase, the total revenue can also decrease. An important takeaway from this exercise is that firms with high marginal costs must be especially careful to temper their enthusiasm for improving operational efficiency.

The impacts of high marginal cost are summarized in Table 7. We first consider scenario 2, which takes the current setting and increases marginal costs. Total costs rise substantially, leading to lower profit, as expected, whereas revenue and consumer surplus are unaffected. Next, consider scenarios 3–7, in which the firm optimizes its prices conditional on service times (similar to scenarios 2–6 in Table 6). When the service times decrease from scenario 5($\tau = [3, 5]$) to scenario 7($\tau = [1, 3]$), we find that the optimal profit drops (from 84.9% to 79.8%), similar to the case with low marginal costs.

However, now we find that the revenue also decreases (from 95.1% to 90.2%) when service time is reduced. With high marginal cost, the firm raises prices substantially, which have particularly strong and negative revenue implications for "smoothing" consumers (high $\alpha_{1,i}$ and low $\alpha_{3,i}$). Specifically, smoothing consumers are much more likely to downgrade or even drop out, leaving only the high bunching consumers (highest $\alpha_{3,i}$), who have much higher valuations for high-end plans when the service time is short. Because the increase in the price is more than offset by the loss of consumers, the firm suffers a net decrease in revenue.

We also evaluate the inefficiency due to the fact that consumers do not pay on the margin. Specifically, we focus on the case whereby the firm can only set the optimal subscription price but fix the per-usage price at the socially efficient level of \$2 each. In this case, there should be no "socially inefficient" consumption, because the consumers bear the full marginal cost of

Table 6. Counterfactual: Improved Service Time

	Scenario					
	1	2	3	4	5	6
Service time	$\tau = [5, 7]$	$\tau = [5, 7]$	$\tau = [4, 6]$	$\tau = [3, 5]$	$\tau = [2, 4]$	$\tau = [1, 3]$
Reduction in τ	0	0	1	2	3	4
Prices	Current	Optimized	Optimized	Optimized	Optimized	Optimized
Revenue (%)	100	96.1	103.5	113.3	122.3	133.4
Cost (%)	100	67.8	76.8	97.2	117.2	154.8
Profit (%)	100	111.8	118.3	122.2	125.2	121.6
Consumer surplus (%)	100	61.6	67.4	82.8	98.7	124.7

Notes. Revenue, cost, and profit are measured as per consumer month. Revenue, cost, and profit at the current service time serve as bases of comparison and are normalized to 100%.

Table 7. Counterfactual: Improved Service Time and Higher Marginal Costs

	Scenario				
	1	2	3	4	5
Service time	$\tau = [5,7]$	$\tau = [4,6]$	$\tau = [3,5]$	$\tau = [2,4]$	$\tau = [1,3]$
Reduction in τ	0	1	2	3	4
Marginal cost	High	High	High	High	High
Prices	Optimized	Optimized	Optimized	Optimized	Optimized
Revenue (%)	100	114.4	107.9	105.4	102.4
Cost (%)	100	124.7	102.7	105.6	111.6
Profit (%)	100	106.8	112.7	106.0	95.6
Consumer surplus (%)	100	109.4	93.6	100.5	97.9

Notes. Revenue, cost, and profit are measured as per consumer month. Optimal quantities at the current service time (and high marginal costs) serve as bases of comparison and are normalized to 100%.

consumption. The results are summarized in Table 8 below. We find that the firm's ability to extract revenue diminishes as the turnaround time becomes shorter, similar to what we had observed earlier. Observe that there is a slight drop of revenue from $\tau = [2,4]$ (117.9% of the current revenue) to $\tau = [1,3]$ (117.1% of the current revenue).

7.4. Mechanism

Overall, these results give us nuanced insights into the boundary conditions at which service can either enhance or hamper revenue and profits. We identify two mechanisms that impact firm outcomes as we reduce the service time:

Mechanism (A): Consumers are able to consume more for two reasons: (a) They are more likely to have more movies in inventory, and (b) waiting cost due to consumption is reduced, because new movies arrive faster. Thus, conditional on the same plan, they consume more, which increases costs. The reduction in service time also increases their valuation for plans, but the difference between the valuation and costs is lower, so less surplus is available for extraction.

Mechanism (B): At long service times, there are two segments of consumers with similarly high valuations: those with high regular consumption utility ($\alpha_{1,i}$), and those with high bunching consumption utility ($\alpha_{3,i}$). As service time increases, both segments increase in average valuation; however, the valuation for bunching consumers increases much more with reduced service time, because they now have access to many more bunching consumption occasions. Observe that these bunching opportunities increase proportionally more than consumption opportunities do. Thus, heterogeneity in valuation can increase when service time decreases, as we also find in Figure 5 below, which shows the density of consumer valuations across plans and service times. In addition, the social surplus at the individual level can also decrease because costs increase at a higher rate than valuation as service time decreases.

Mechanism (B) examines the role of heterogeneity in valuations for the plans, and how that impacts the firm's ability to extract surplus.

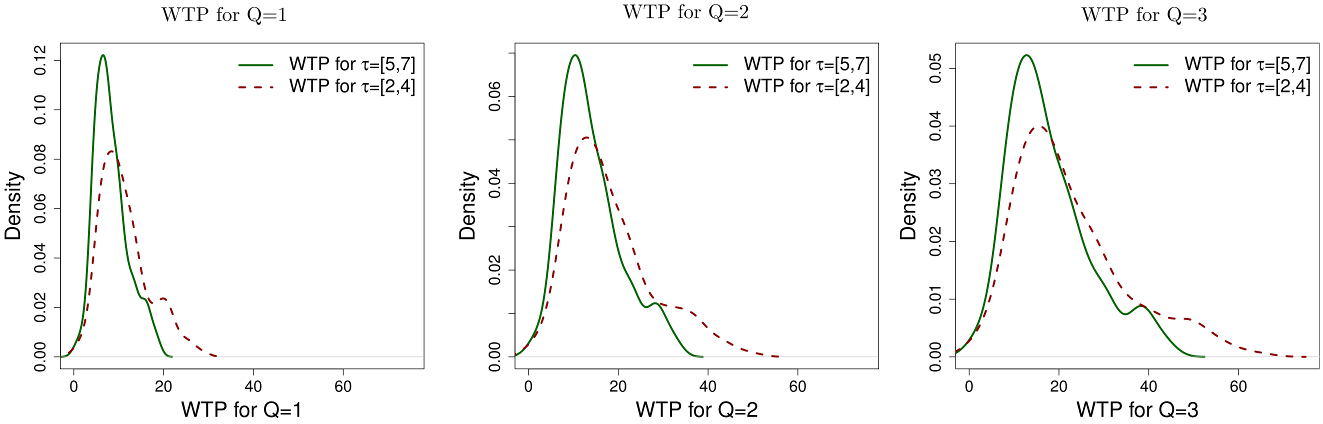
First, we find that improved service time increases consumer valuations as expected. However, it also increases the variance of these valuations *within each plan*, which can make surplus extraction more challenging. Figure 5 shows how the willingness to pay changes, and the summary statistics of the WTP distribution are presented in Table 9. We find that as service time decreases, the mean valuation increases but the variance in valuation also increases. Increased variance in valuations makes it more challenging to extract surplus.[14]

Table 8. Counterfactual: Optimal Subscription with the Socially Efficient Unit Price

	Scenario					
	1	2	3	4	5	6
Turnaround time	$\tau = [5,7]$	$\tau = [5,7]$	$\tau = [4,6]$	$\tau = [3,5]$	$\tau = [2,4]$	$\tau = [1,3]$
Reduction in τ	0	0	1	2	3	4
Prices (%)	Current	Optimized	Optimized	Optimized	Optimized	Optimized
Revenue (%)	100	92.1	96.7	102.0	117.9	117.1
Cost (%)	100	71.8	78.3	92.9	118.5	129.7
Profit (%)	100	103.1	106.7	106.8	117.4	110.0
Consumer surplus (%)	100	72.7	75.5	86.9	100.4	113.9

Notes. Revenue, cost, and profit are measured as per consumer month. Revenue, cost, and profit at the current turnaround time serve as the bases of comparison and are normalized to 100%.

Figure 5. Changes in Consumer Valuation with Reduced Service Time



Second, improved service time also makes segmentation difficult because the variation in valuation *across plans decreases*. Figure 6 shows the overlap in valuations across the three plans $Q = 1,2,3$. As the service time is reduced, the overlap between the plans increases. For example, the overlap between $Q = 1$ and $Q = 2$ goes from 46% under worse (high) service time to 51% under better (low) service time. Because of this increased overlap, the plans are not able to effectively segment consumer valuations. In other words, consumers see the plans as closer substitutes when the service time is reduced. This results in a greater difficultly in segmentation of consumers by using the three plans, which in turn makes surplus extraction more challenging for the firm.

Overall, owing to both of the above effects, it becomes more difficult to extract surplus, because of *increased* consumer heterogeneity in valuations for each plan and *reduced* consumer heterogeneity for variation across plans.

7.5. Alternative Pricing Strategies

In the multidimensional screening literature focusing on pricing, most studies focus on optimizing prices given a fixed pricing mechanism. However, we have seen above that such an approach might not be sufficient and that profits might decline as service time improves (decreases). Given potential misalignment between the firm's current subscription pricing strategy with fast service, we explore other pricing strategies.

7.5.1. Pricing Based on Service Time. First, consumers with higher service time derive less value from the service *and* are less costly for the firm to serve, compared with consumers with lower service time. Informed by the literature on spatial price discrimination (Miller and Osborne 2014, Ngwe 2017), we allow the firm to charge different prices according to service time. Thus, the price is specified as $\boldsymbol{P}$ for consumers with lower service time and $\lambda(\tau)P$ for consumers with higher service time. In this counterfactual, the firm again maximizes the expected total profits:

$$\max_{\boldsymbol{P},\lambda(\tau)} \sum_{i=1}^{200} \sum_{m=1}^{M} \sum_{k=1}^{3} \beta^{mT} \times Prob(q_{ik}|\boldsymbol{P}, \lambda(\tau), \tau_i) \times \left[\left(p(q_{ik}|\boldsymbol{P}, \lambda(\tau), \tau_i) - mc \times \sum_{t=1}^{T} E(c_{it}|q_{ik}, \tau_i) \right) \right]. \quad (8)$$

In Equation (8), subscription prices are based on service time. We find that the firm optimally lowers the price for consumers with higher service time. Offering customized subscription prices can increase the firm's profitability, compared with the current case in which the same subscription prices are offered to consumers. However, a caveat is in order: firms may face a strong negative consumer reaction from doing such price discrimination (e.g., Amazon had to reverse its decision for this reason).

7.5.2. Pricing per Unit Consumption—Two-Part Tariff. With a subscription plan, consumers do not face consumption costs, and under short service time they might be induced to consume "excessively," even in cases when consumption value does not exceed the marginal cost. Thus, we examine including a per-unit

Table 9. Consumer Valuation for Plans under Long and Short Service Times

	Service time	
Plan	Long, $\tau = 5,7$	Short, $\tau = 2,4$
$Q = 1$	8.28 (3.72)	11.26 (5.52)
$Q = 2$	13.85 (6.74)	18.12 (9.56)
$Q = 3$	17.76 (9.20)	22.50 (12.55)

Note. Values are mean (standard deviation) of valuation ($).

Table 10. Counterfactual: Price Segmentation by Service Time

	Scenario					
	1	2	3	4	5	6
Service time	$\tau = [5,7]$	$\tau = [5,7]$	$\tau = [4,6]$	$\tau = [3,5]$	$\tau = [2,4]$	$\tau = [1,3]$
Reduction in τ	0	0	1	2	3	4
Prices	Current	Optimized	Optimized	Optimized	Optimized	Optimized
Revenue (%)	100	120.4	123.7	131.1	142.1	148.4
Cost (%)	100	95.8	97.3	108.7	134.3	150.0
Profit (%)	100	132.9	138.6	142.4	145.8	147.1
Consumer surplus (%)	100	72.0	70.4	76.2	94.4	106.1

Notes. Revenue, cost, and profit are measured as per consumer month. Revenue, cost, and profit at the current service time serve as bases of comparison and are normalized to 100%.

price for each consumption, in addition to the subscription price. The firm's problem is to maximize over subscription price $\boldsymbol{P}$ and unit price ρ:

$$\max_{\boldsymbol{P},\rho} \sum_{i=1}^{200} \sum_{m=1}^{M} \sum_{k=1}^{3} \beta^{mT} \times Prob(q_{ik}|\boldsymbol{P}, \rho) \times \left[\left(p_k - (mc - \rho) \times \sum_{t=1}^{T} E(c_{it}|q_{ik}, \rho, \tau_i) \right) \right]. \tag{9}$$

Both purchase and consumption decisions are impacted by the marginal price, ρ. We find that the two-part tariff improves the firm's profitability (Section 7.5.2), primarily resulting from lower costs.

We note that "unlimited" is often a popular catch phrase advocated by various subscription services with products ranging from movies, games, and books to mobile phone plans. However, allowing consumers free access might be problematic for two reasons. First, the firm's ability to extract value from incremental products might not be commensurate with the costs of honoring such a commitment. Further, adverse selection is likely when heavy users find the unlimited policy more attractive. Broadly speaking, this problem is endemic to similar "all-you-can-eat" services and needs to be addressed when marginal costs are significant.

We observe a few different ways this issue is addressed in practice. Amazon and Staples both offer a free-shipping policy but impose minimum spending amounts of $25 and $49.99, respectively. In the context of online movie rental, the problem became well-known after the *throttling* approach taken by Netflix, which targeted its heavy users.[15] Our findings provide clear empirical support for this practice, especially when service time has improved significantly. More broadly, we present a set of pricing strategy options to understand which ones can serve to alleviate this problem.

7.6. Digital Delivery (Online Streaming)

Although many RBM services (e.g., designer dresses and baby toys) will continue to rely on mail or courier services, information goods such as movies are increasingly delivered online thanks to developments in streaming technology.[16]

Migration to online delivery has two effects for the firm based on our model. First, waiting time for the movies is reduced to zero, and willingness to pay increases. Second, it could also leads to a significant change in the cost structure of the firm. We note that both marginal costs and fixed cost-based licensing are commonly used in practice, with newer movies typically being licensed per viewing.[17]

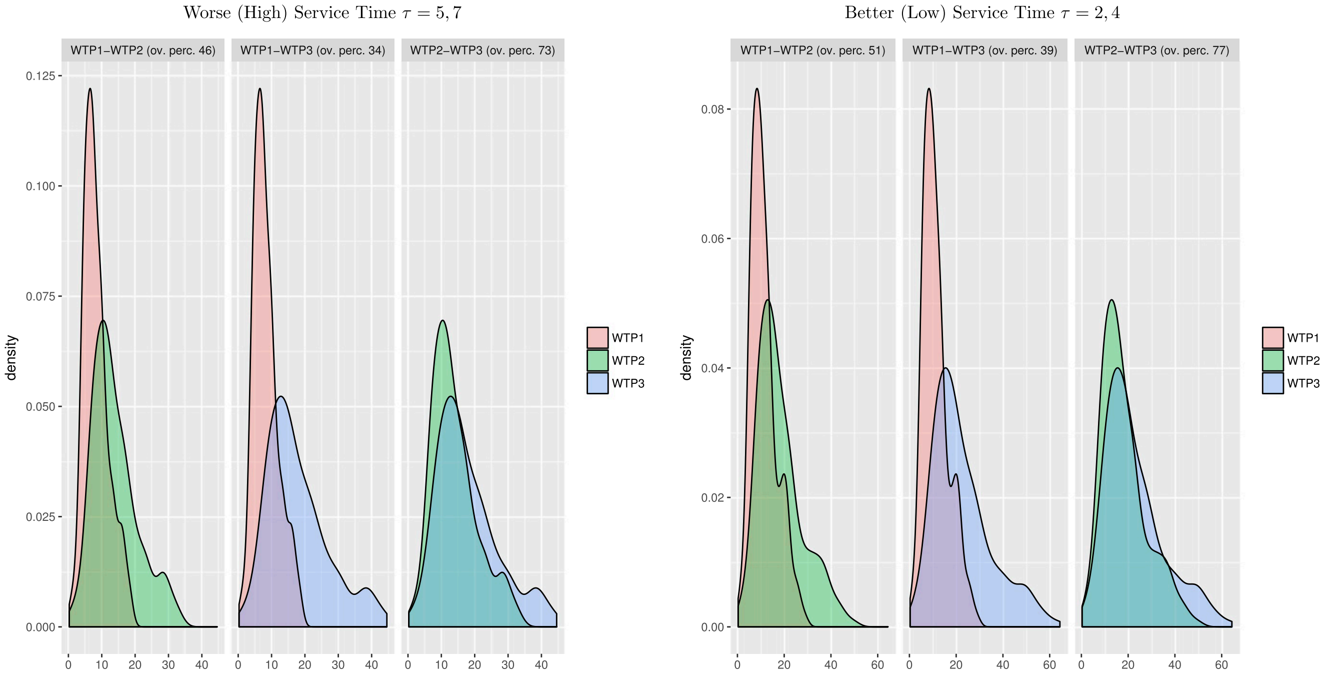
We observe interesting differences in the choice of price formats. Netflix chooses a subscription model but consolidated the original multitier subscription-price format with a simple "all-you-can-watch" streaming

Table 11. Counterfactual: Optimal Subscription Price and Optimal Unit Prices

	Scenario					
	1	2	3	4	5	6
Service time	$\tau = [5,7]$	$\tau = [5,7]$	$\tau = [4,6]$	$\tau = [3,5]$	$\tau = [2,4]$	$\tau = [1,3]$
Reduction in τ	0	0	1	2	3	4
Prices (%)	Current	Optimized	Optimized	Optimized	Optimized	Optimized
Revenue (%)	100	93.9	106.1	115.8	124.4	126.1
Cost (%)	100	55.3	69.6	93.3	105.3	115.2
Profit (%)	100	114.9	126.0	128.0	134.7	131.9
Consumer surplus (%)	100	52.1	56.3	73.7	78.9	88.5

Notes. Revenue, cost, and profit are measured as per consumer month. Revenue, cost, and profit at the current service time serve as bases of comparison and are normalized to 100%.

Figure 6. Difference in Plan Valuations Across Service Times



service, allowing consumers to access unlimited streaming for a fixed subscription fee per month. iTunes, another major online streaming provider, chooses à la carte pricing, whereby the consumer pays approximately $5–$7 to watch a movie within a 24–48 hour period.

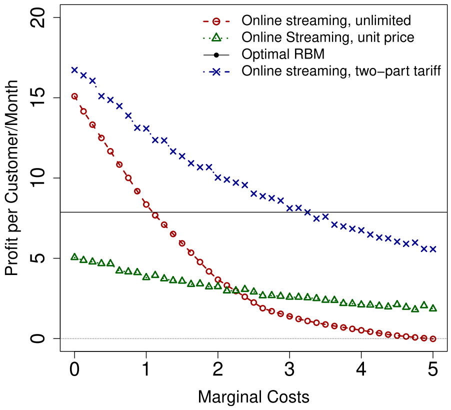
We conduct an illustrative exercise to identify the revenue impact on our focal firm across different pricing formats. If the managers of our focal firm were considering the revenue impacts of online streaming when they do not have any streaming data, this would provide one of the only ways to obtain an estimate. Moreover, our interest is in understanding the underlying mechanism of how consumer and firm decisions change and serve to explain observed pricing practices.

However, there are a number of assumptions required in using consumer preferences that we have recovered for watching physical DVDs to evaluate their utility for streaming. We list below some of the assumptions and caution the reader that this counterfactual or extrapolation is likely to depend significantly on the nature and validity of these assumptions. First, video quality may be higher or lower with streaming compared with DVDs/Blu-Ray media. Second, consumers might value streaming for convenience in streaming across a range of devices, as well as for its ease of use. Third, search processes may be easier with streaming, although even with the DVD format, consumers had access to movie information, trailers, and consumer reviews. Broadly, the above factors increase or decrease consumer valuation or willingness to pay for the streaming service, and examining these valuations would provide a more comprehensive idea.

We assume the firm does not have cost data yet, so we examine a wide range of cost levels. We focus on the case in which the firm faces a marginal cost-based licensing fee for each movie viewed. We consider monthly subscription prices (e.g., Netflix) and per-unit prices (e.g., iTunes) for the firm and compute profits for each price format at marginal cost between $0 and $5.[18] The short-run intertemporal consumption trade-offs created by service time is not present with streaming. With à la carte pricing, daily consumption is determined as the current consumption that gives the consumer highest net utility. We check to ensure that the optimality is obtained at interior price points. The results are summarized in Figure 7.

Figure 7 compares the profits from the optimal subscription pricing (red circles) and the optimal à la carte (green triangles), the profit (dollars per consumer month) as the outcome of interest. We find that subscription pricing dominates à la carte pricing if the marginal cost is low (less than 0.2); and the reverse is true if the marginal cost is high (more than 0.2). Intuitively, a high marginal cost should induce the firm to price on the basis of usage, to limit inefficient consumption (i.e., the consumers' valuation of the content is low, compared with the marginal cost of providing it). In contrast, a low marginal cost implies that the firm should not restrict consumption, which helps the firm to charge higher subscription prices. This comparison offers a likely explanation as to why "unlimited" streaming services that charge a monthly price (e.g., Netflix) do not offer the latest movies

Figure 7. Comparisons Between Unlimited and à la Carte Streaming Services



and more expensive content, despite a very large consumer demand for such content. Instead, Netflix has focused on providing content with low marginal costs of licensing, including original content for which it likely has to pay large fixed costs. In contrast, iTunes, which does provide the latest movies, does not offer a subscription, and each movie has a per-unit price.

Comparing profits from streaming with optimal RBM shows a surprising result: firm profitability is lower for streaming services with either subscription pricing or à la carte pricing. In other words, the firm could seemingly be *harmed* by the technology change. In particular, streaming with subscription is not desirable even if it is technologically feasible to do so, unless the marginal cost is sufficiently low. This happens because with the unlimited subscription plan, the firm has not been able to price discriminate across consumers on the basis of quantity or another metric. However, it must be noted that there may be other ways to price discriminate that we have not considered here, either according to quality (resolution of content) or according to the number of screens. We see such approaches used in practice and note that these might be required in the streaming case because it becomes challenging to price discriminate on the basis of quantity with a monthly plan.

We further consider a possible two-part tariff, which combines the subscription and a per-usage fee. In this scenario we search for both types of optimal prices simultaneously. The resulting profit is demonstrated with the blue line in Figure 7. We find that the optimal two-part tariff is expected to generate higher profits than both subscription pricing and à la carte. Importantly, optimal the two-part tariff is more profitable than RBM pricing at the current marginal cost of $2. Overall, switching to digital delivery has profit potential for the firm, but only if the firm can effectively manage its marginal costs (e.g., licensing fees) and choose the appropriate price format.

It is important to note that the above results are only based on the model developed in this paper and must be interpreted with caution. There are a number of factors that we have not modeled that we expect would be important in practice. For example, online distribution can create significant market expansion effects and reach new markets beyond the customers currently served. If the firm caters to a niche segment of customers focused on differentiated content, it is more likely that such expansion effects might be small, whereas if the firm has a large market potential, then market expansion effects may be large and easily outweigh the profit loss that we find.

8. Discussion and Conclusion

We examine the effectiveness of second-degree pricing strategies, where consumers self-select a plan, under the conditions of operational and technological transformation that result in different service time, which can be viewed as a dimension of quality.

We develop a framework for understanding forward-looking consumers' decisions in our empirical context, involving a closed-loop rental-by-mail firm. Consumers make short-run consumption choices, and we find significant heterogeneity in how consumers make intertemporal trade-offs. Some consumers prefer to smooth their consumption over time, whereas others prefer to bunch their consumption, which might lead to different implications for service time. For long-run purchase decisions, we develop a structural model of ordered choice, which can be applied and adapted to a number of settings where choices have a clear vertical structure, with smaller plans nested in larger ones. We would find such similar plans whenever firms offer different plan versions (e.g., with cable or internet services such as Hulu or ESPN), or freemium models, or mobile phone service provider plans specifying the amount of data.

We find significant heterogeneity in consumer intertemporal preferences, with a majority smoothing their consumption and a significant minority bunching. As movie watching adapts to online streaming, binge watching or bunching is also enabled by design features; for example, streaming services suggest new episodes of shows or auto-play them, and the content is designed to leverage bunching.

We find that firm profits can be nonmonotonic in terms of service time, owing to two separate mechanisms. First, in mechanism (A), as the service time increases, consumers' wait time is shorter, and there are fewer occasions when they are stocked out. This increases the consumption probability within each plan and therefore increases the costs. The difference between willingness to pay and costs might decrease for some consumers.

More importantly, we demonstrate a novel mechanism (B), which to our knowledge has not been suggested in prior literature. We find that improved (reduced) service time increases the consumer heterogeneity in valuation of the firm's product (plans), thus making it more challenging to extract the surplus as revenue. Thus, there is a distinct divergence between value creation for consumers and value capture by the firm. We find two reasons for this divergence. First, through better service time, the firm might create increased value for each of its customers, but consumer valuation within each product (plan) becomes more heterogeneous, reducing its ability to capture surplus. Second, as service time improves, the valuation across plans becomes more similar, implying that segmentation using multiple products is not as effective. Thus, each effect reinforces the challenge for the firm in capturing a proportionate share of created surplus.

Taken broadly, our findings suggest that an improvement in one functional area (operations) might lead to diminished ability in another area (marketing), underlining the challenges in coordinating both choices into the firm's overall strategy and chosen business model.

Our current research has specific limitations, which in turn open up important avenues for future research. First, our results are based on a specific rental service, and it would be helpful to empirically examine other settings. In particular, although the utility from movie consumption is generally derived within a day, other products (e.g., video games and designer dresses) may give consumers a stream of utilities over a longer period. Future research can readily adapt our modeling framework to those services. From a modeling perspective, it might be useful to understand whether consumers can transition between latent states, such as in a hidden Markov model (Netzer et al. 2008), which would likely require more plan switches than are available in our setting. Within the RBM setting, we have abstracted away from several issues. First is the potential preference heterogeneity over products or content, which is perhaps less important in our empirical setting but might well be critical in other settings. Second, and relatedly, we could examine how the movie queue dynamics work and how consumers choose to add titles to queues.

More broadly, beyond the RBM context, it would be helpful to understand how other firms' strategic choices impact consumer heterogeneity and selection, and through this, the mechanisms of surplus creation and extraction, which impact firm profitability.

Acknowledgments

The authors thank attendees of seminars at CKGSB, Columbia University, Duke University, Emory University, Peking University, Shanghai University of Finance and Economics, Tsinghua University, and University of Texas at Dallas for their feedback. The authors contributed equally and are listed in alphabetical order.

Appendix A. The Rental-by-Mail Model

Figure A.1. Illustration of the RBM Model

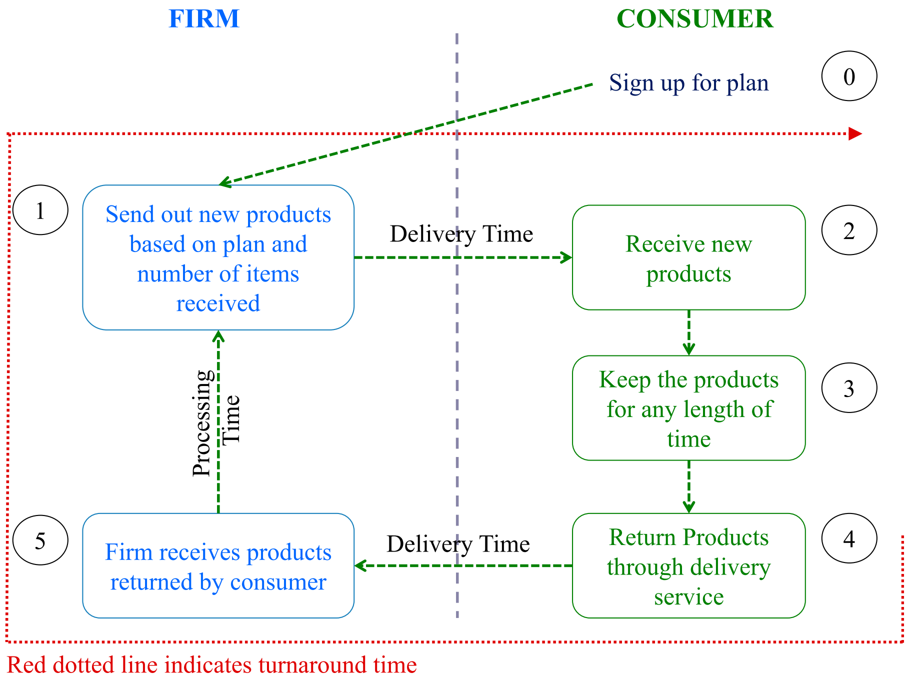

FIRM
CONSUMER
Sign up for plan
0
1
Send out new products based on plan and number of items received
Delivery Time
Receive new products
2
Keep the products for any length of time
3
Processing Time
5
Firm receives products returned by consumer
Delivery Time
Return Products through delivery service
4
Red dotted line indicates turnaround time

Table A.1. Plans Offered by Representative RBM Services

	Delivery	Product selection	Subscription plan	Fee
Netflix	Mail	100,000 + movies	1 at a time	$7.99 per month
			2 at a time	$11.99 per month
			3 at a time	$19.99 per month
GameFly	Mail	8,000 + games and movies	1 at a time	$10 per 2 months
			2 at a time	$20 per 2 months
Rent the Runway	UPS or courier services	1,000 + dresses	3 at a time	$99 per month
TurningArt	Mail	1,000 + works of art	1 at a time	$15 per month
			2 at a time	$20 per month
BookFree	Mail	250,000+ titles	2 at a time	$8 per month
			4 at a time	$10 per month
			6 at a time	$13.50 per month
			9 at a time	$18 per month
			12 at a time	$22.25 per month
			15 at a time	$27.50 per month

The RBM model is used for a wide array of rental products, including movies (e.g., Netflix), books (e.g., BooksFree), art works (e.g., TurningArt), video games (e.g., GameFly), and toys (e.g., BabyPlays). Consumers have the convenience of receiving and returning rental products in the mail (versus traveling to physical stores) and a deeper selection of rental products,[19] with the trade-off being a delay in obtaining the product.

Figure A.1 details the process dynamics in the RBM model, whereby the firm only ships a "new" product when it receives a product returned by the consumer. The *service time* is the time interval between step 4 (when the consumer sends out consumed products) and step 2 (when the consumer receives new products). Thus, although the RBM service gives consumers the convenience of receiving products at home, it also implicitly limits the number of movies they can watch in a subscription cycle according to the service time, which separates the sequence of consumption opportunities for the consumer.

Table A.1 details examples of firms using the RBM model.

Appendix B. Ordered Model for the Consumption Decision

We consider how to obtain choice probabilities for the ordered choice model. Table B.1 details the notation.

Let $u(c,s;\nu)$ be the utility corresponding to the consumption choice. We next derive the thresholds and choice probabilities. The ex ante short-run value function is denoted V, and because the state transition is deterministic in the short run, we write $V(s'|s,c)$.

$$u(c,s,\nu) = \alpha_1 c + \alpha_2 c^2 + \beta V(s'|s,c) + \underbrace{\alpha_3 c \quad \nu}_{\text{Stochastic Consumption}}$$

The choice probabilities are generated from the shock ν, which we assume is distributed as LogNormal(0,1). Any continuous distribution with positive support would work in its place.

Observe that when $\alpha_3 > 0$, a higher shock generates higher utility from consuming more, leading to higher consumption choices. Thus, the optimal consumption $c^* \in \{0,1,\ldots,Q\}$ (where Q is the quota of the consumer) weakly increases in the shock ν (for $\alpha_3 > 0$). Below we characterize thresholds that are then used to define choice probabilities.

B.1. Characterizing Thresholds and Choice Probabilities

To characterize the regions corresponding to the optimal consumption choices, we define thresholds τ_{jk} that determine how the consumer will make different choices k and j. A consumer prefers to choose $c_k > c_j$ if $\nu < \tau_{kj}$, and $c_j > c_k$ if $\nu \geq \tau_{kj}$. The thresholds for $0 < k < j \leq Q$ can be derived from the above utility specification as

$$\tau_{kj}(s;\alpha) \overset{def}{=} \frac{\alpha_1(c_j - c_k) + \alpha_2\left(c_j^2 - c_k^2\right) + \beta\left[V\left(s'|s,j\right) - V\left(s'|s,k\right)\right]}{\alpha_3(c_k - c_j)}.$$

Observe that the thresholds must be defined for each short-run state and are also dependent on the short-run value function V. Next, choice probabilities can be obtained from the distribution of the shock ν(cdf F_ν) as follows:

$$P(c^* = j) = \begin{cases} F_\nu(\tau_{0,1}) & j = 0 \\ F_\nu(\tau_{j,j+1}) - F_\nu(\tau_{j-1,j}) & 0 < j < Q \\ 1 - F_\nu\left(\tau_{(j-1),j}\right) & j = Q. \end{cases}$$

The ordered choice model for plan choices is derived in a similar manner, obtaining choice probabilities from thresholds.

Table B.1. Ordered Choice Model Notation

Notation	Interpretation
c_{it}	Individual i 's consumption choice at time t, $c_{it} = 0,1,2,3$
Q_{it}	Individual i 's quota choice at time t, $Q_{it} = 0,1,2,3$
$V\left(s'\|s,c\right)$	Short-run value function at short-run state s and consumption c
α_p	Price coefficient
α_{sw}	Switching cost
α_8	Importance of plan-level utility shock
τ_{ij}	Threshold between consumption levels i and j

Appendix C. Model and Estimation Details

C.1. Transition of the Mailing State

The state x_{it} evolves to $x_{i,t+1}$, on the basis of the following law of motion when there is no change in plan:

$$\begin{aligned} x^0_{i,t+1} &= x^0_{it} + x^1_{it} - c_{it} \\ x^k_{i,t+1} &= x^{k+1}_{it}, \qquad 1 \le k \le (\tau - 1) \\ x^\tau_{i,t+1} &= c_{it}. \end{aligned} \tag{C.1}$$

The first line of Equation (C.1) states that the inventory $x^0_{i,t+1}$ that will be available to consumer i at $t+1$ is the current inventory (x^0_{it}), plus any movies that she will receive (x^1_{it}), less her current period consumption (c_{it}), which is shipped back to the firm. The last line of Equation (C.1) means that after the consumer returns the just-watched movies to the firm, she will receive the same number of movies, but only after the full service time of τ days. The middle $(\tau - 1)$ lines of Equation (C.1) have a straightforward interpretation: on day $t+1$, movies in the mail are one day closer to being delivered to the consumer. Note that $\sum_{k=0}^{\tau} x^k_{i,t+1} = \sum_{k=0}^{\tau} x^k_{it} = q_{it}$: in the closed loop RBM rental process, the total number of movies in the mail, plus the consumer inventory, is always equal to the quota on any given day.

To illustrate the dynamics of the mailing states, consider consumer i in Table C.1. The consumer has a service time of $\tau = 5$ days and subscribes to a plan with three movies (i.e., $q = 3$). Suppose that the initial mailing state for the consumer is $x_{it} = (2, 0, 0, 0, 1, 0)$, so that she currently holds two movies and expects one movie to arrive in four days. The rows of Table C.1 illustrate how the state $x_{i,t+1}$ evolves when the consumer chooses either not to watch any movies in the current period (i.e., $c_{it} = 0$) or watch one ($c_{it} = 1$) or two ($c_{it} = 2$) movies. For example, if $c_{it} = 0$ (no consumption), the inventory available to consumer i remains unchanged at two on the next day. Meanwhile, she is one day closer to receiving a movie in the mail ($x^3_{i,t+1} = 1$ and $x^4_{i,t+1} = 0$, versus $x^3_{it} = 0$ and $x^4_{it} = 1$).

Table C.1. Illustration of the Mailing State Transition for $\tau = 5$ Days.

State	x^0_{it}(inventory)	x^1_{it}	x^2_{it}	x^3_{it}	x^4_{it}	x^5_{it}
Initial state x_{it}	**2**	**0**	**0**	**0**	**1**	**0**
$x_{i,t+1}$ with $c_{it} = 0$	2	0	0	1	0	0
$x_{i,t+1}$ with $c_{it} = 1$	1	0	0	1	0	1
$x_{i,t+1}$ with $c_{it} = 2$	0	0	0	1	0	2
$x_{i,t+1}$ with $c_{it} > 2$ (infeasible)	—	—	—	—	—	—

C.2. Transition of the Weekend State

Formally, the transition process for the day of the week is specified as

$$w_{i,t+1} = \begin{cases} w_{it} + 1, & w_{it} < 7 \\ 1 & w_{it} = 7. \end{cases} \tag{C.2}$$

The intraweek dynamics are driven by the exogenous state of weekends versus weekdays. If consumers have a different (e.g., higher) utility for weekend consumption, then we would expect higher consumption during the weekends, compared with weekdays, for two reasons. First, the effect of higher instantaneous utility during the weekends would lead to higher consumption probability, conditional on all other state variables. Second, consumers would reduce their consumption on weekdays to ensure that they have sufficient movies available to watch during the weekend, which is also impacted by the service time.

C.3 Estimation of Ω

We estimate separately from the data in a first stage. We first discretize the data into $\omega = 3$ bins and estimate the $(N_\omega \times N_\omega)$ transition matrix, imposing the following restrictions. Because the content set only evolves as an increasing process, we set the nondiagonal elements of the lower triangular matrix to zero. We also allow increasing transitions only to the next higher state for simplicity and because the data in the content set support this transition.

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} & 0 \\ 0 & \Omega_{22} & \Omega_{23} \\ 0 & 0 & \Omega_{33} \end{bmatrix} \tag{C.3}$$

We estimate the parameters $\Omega_e = (\Omega_{11}, \Omega_{12}, \Omega_{22}, \Omega_{23})$ nonparametrically using a bin estimator. Note that $\Omega_{33} = 1$ is fixed, because there is no other state to which it can transition.

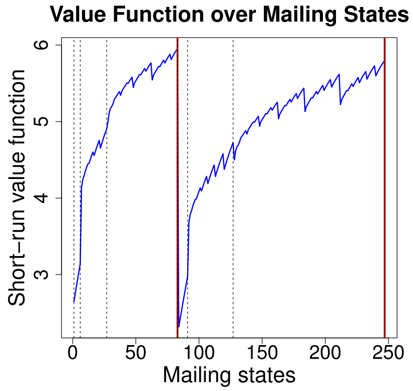
C.4. Value Function over Mailing States

We plot the value function in Figure C.1 for a representative consumer (whose preference parameters are fixed at the posterior means of the individual parameters) at the first day of the payment cycle. There are two subspaces, corresponding to the two levels of service time: states to the left (right) of the first red solid line correspond to $T = 5$ (7) days, respectively. The dashed black lines separate the states for the three subscription plans. Observe that the consumers' value functions are higher for the shorter service time, reflecting the fact that the consumers derive less value from a longer service time. Figure C.1 clearly demonstrates that the shape of the value function is jagged. The mailing states for the different service times are disjoint and can be separated in the estimated process.

C.5. Outline of the MCMC Algorithm

Below, we outline the steps of the MCMC algorithm that generates the posteriors of the parameters Θ in the utility

Figure C.1. Value Function



function: $\Theta = (\alpha_{1,i}, \alpha_{3,i}, \alpha_w, \alpha_{cs}, \alpha_p, \alpha_{sw})$, as well as the hyperparameters that govern the distribution of the heterogeneous parameters: $\Xi = (\Delta, V_\alpha)$. Table C.2 provides further details.

Estimation for the structural parameters is done using the IJC method of the Bayesian estimation of dynamic discrete choice models developed by Imai et al. (2009). The estimation was implemented in *Julia*. In the baseline specification, *for identification,* $\alpha_{2,i} = -1$ *is fixed for all consumers throughout the entire MCMC process.* For the semiparametric model, *similarly* $\alpha_{3,i} = +1$ *is fixed, and the heterogeneous parameters are* $(\alpha_{1,i}, \alpha_{2,i})$.

The estimation steps outlined below provides a flexible process for the estimation procedure of both homogeneous and heterogeneous parameters.

The MCMC process has three blocks: Block I (Step 1) draws the set of $NPAR_{homo}$ homogeneous parameters $\gamma = (\alpha_w, \alpha_{cs}, \alpha_p, \alpha_{sw})$. Block II (Step 2) draws the set of $NPAR_{hetero}$ heterogeneous parameters, $\alpha_i = (\alpha_{1,i}, \alpha_{3,i})$, $\forall i$. Block III (Steps 3 and 4) draws the hyper-parameters Ξ.

Step 0

At the beginning of each iteration r, start with the history $\mathcal{H}^r$:

$$\mathcal{H}^r = \left\{(\gamma^*, \alpha_i^*)^l, E\tilde{V}\left(s, \gamma^{*l}, \alpha_i^{*l}\right)_{l=r-N}^{r-1}, q(\gamma^{r-1}, \alpha_i^{r-1}|DATA_i)_{i=1}^{I}\right\}, \tag{C.4}$$

where I is the total number of consumers in the estimation sample, and N is the number of past iterations that will be used for the approximation of the Emax function in Steps 1 and 2. The IJC algorithm makes an important distinction between the history of candidate parameters and accepted parameters. We use plain subscript (e.g., $r-1$) to denote the previously accepted parameters and use superscript $*$ to denote the candidate parameters.

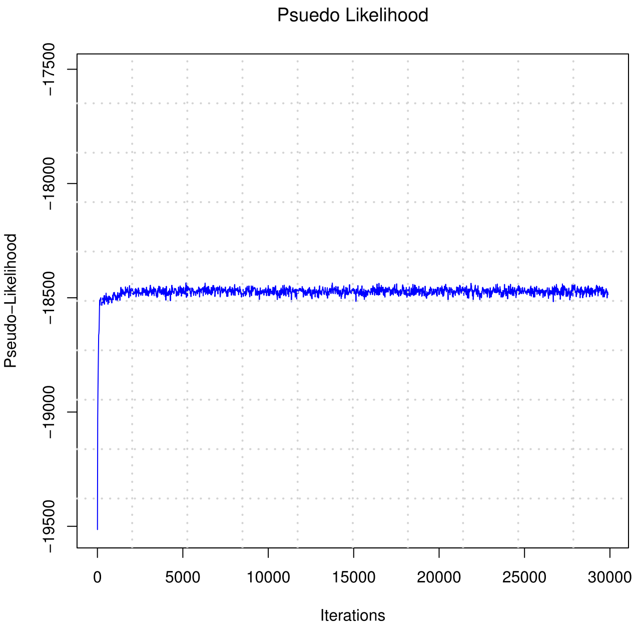
It is important to note that the approximation is only for the long-run value function. The short-run value functions are finite-horizon and computed exactly for each parameter value.

The first component of the history $(\gamma^*, \alpha_i^*)^l = (\alpha_{w*}, \alpha_{cs*}, \alpha_{p*}, \alpha_{sw*}, \alpha_{1,i}^*, \alpha_{3,i}^*,)^l$ is the vector of candidate parameters from the past iteration l. The second component, $E\tilde{V}(s, \gamma^{*l}, \alpha_i^{*l})$, is the corresponding pseudo-value function. Conceptually, the pseudo-value function is an approximation to the full solution of the true value function: the sequence of pseudo-value function converges to the true value function in probability uniformly (Imai et al. 2009). The third and final component, $q(\gamma^{r-1}, \alpha_i^{r-1}|DATA_i)$, is the pseudo-likelihood function, which depends on the previously accepted parameters $(\gamma^{r-1}, \alpha_i^{r-1})$. Apparently, $q(\gamma^{r-1}, \alpha_i^{r-1}|DATA_i)$ also depend on $DATA_i$, which is the sequence of the observed consumption and purchase decisions $\{d_{it}\}_{t=1}^{T_i}$ plus the payoff relevant state variables (e.g., sequence of DVDs, previous plan choice, weekend, day of the month, and DVD content size) at time t. To reduce clutter in the notation, we suppress the dependence of q on $DATA_i$ and write $q(\gamma^{r-1}, \alpha_i^{r-1})$ instead.

Table C.2. Estimation Details

Estimation component	Value
No. of pseudo-value functions in history	2,400
No. of individuals	200
Overall homogeneous acceptance rate (%)	32.2
Overall heterogeneous acceptance rate (%)	26.3
Bandwidth for kernel	Based on Silverman (1986)

Figure C.2. Convergence of Pseudo-likelihood



Step 1. In this step, we draw the homogeneous parameters γ^r using the random-walk Metropolis-Hastings algorithm.

1a. On the basis of the current draw of γ^{r-1}, use a multivariate normal density (i.e., $N(\gamma^{r-1}, \Sigma)$) as the proposal density. Σ is an $NPAR_{homo}$ by $NPAR_{homo}$ diagonal matrix whose elements are proportional to the parameter estimates of homogeneous parameters obtained from the maximum likelihood estimation. The vector of candidate parameters is denoted γ^{r*}.

1b. For each consumer, compute the pseudo-likelihood at the candidate vector γ^{r*}, that is, $q(\gamma^{r*}, \alpha_i^{r-1})$, conditional on $DATA_i$, and $E\hat{V}(s, \gamma^{r*}, \alpha_i^{r-1})$, which is the Emax function approximated by the weighted average of the N past pseudo-value functions:

$$E\hat{V}(s, \gamma^{r*}, \alpha_i^{r-1}) = \sum_{l=r-N}^{r-1}\left[E\tilde{V}\left(s, \gamma^{l*}, \alpha_i^{r-1}\right)\right] \cdot \frac{K_h(\gamma^{r*} - \gamma^{l*}, \alpha_i^{r-1} - \alpha_i^{l*})}{\sum_{k=r-N}^{r-1} K_h(\gamma^{r*} - \gamma^{k*}, \alpha_i^{r-1} - \alpha_i^{k*})}. \tag{C.5}$$

Notice that the pseudo-value functions, which are the bases of the nonparametric approximation, correspond to the previous candidate parameters, which can have larger variations compared with the accepted parameters. The weights for each of the N pseudo-value functions are based on the distances between the candidate vector and the previously stored vectors γ^l. We use the standard Gaussian kernel for the nonparametric approximation:

$$K_h(\gamma^{r*} - \gamma^{l*}, \alpha_i^{r-1} - \alpha_i^{l*}) = \frac{2\pi}{NPAR} exp\left[\sum_{j=1}^{NPAR_{Homo}} \frac{(\gamma^{r*} - \gamma^{l*})^2}{2h_j} + \sum_{j=1}^{NPAR_{Hetero}} \frac{(\alpha_i^{r-1} - \alpha_i^{l*})^2}{2h_j}\right], \tag{C.6}$$

where $NPAR$ is the dimension of Θ, and $NPAR = NPAR_{homo} + NPAR_{hetero}$. We use h_j to denote the bandwidth or smoothing parameter for the jth parameter. The selection of the bandwidth is based on the trade-off between the bias and variance of the resulting estimator. Using Silverman's rule of thumb (Silverman 1986), we set $h_j = \hat{\sigma}_j N^{-\frac{1}{5}}$, where $\hat{\sigma}_j$ is the sample standard deviation of N sample points.

1c. Similarly, we compute the pseudo-likelihood $q(\gamma^{r-1}, \alpha_i^{r-1})$ at the previously accepted vector γ^{r-1}:

Assuming a diffuse prior for γ, we determine whether to accept γ^r based on the following acceptance probability:

$$Prob_{accept} = \min\left[\frac{\prod_{i=1}^{I} q(\gamma^{r*}, \alpha_i^{r-1})}{\prod_{i=1}^{I} q(\gamma^{r-1}, \alpha_i^{r-1})}, 1\right]. \tag{C.7}$$

If accept, set $\gamma^r = \gamma^{r*}$; if reject, set $\gamma^r = \gamma^{r-1}$.

Step 2. In this step, we use the random-walk Metropolis-Hastings algorithm to draw α_i^r for each consumer.

2a. α_i is distributed as $N(\Delta^r Z_i, V_\alpha)$. In this general specification, Z_i can include time-invariant demographic variables (e.g., family size), so that we can relate the systematic differences in $(\alpha_{1,i}, \alpha_{3,i})$ to these demographic differences. For example, it is possible that consumers who have larger family size may have higher baseline consumption utilities $(\alpha_{1,i})$. Unfortunately, the focal firm does not have such information available. So in practice, Z_i is a vector of ones in our model specification.

We first generate a candidate α_i^{r*} as $\alpha_i^{r*} \sim N(\alpha_i^{r-1}, \Psi)$, where Ψ is an $NPAR_{hetero}$ by $NPAR_{hetero}$ diagonal matrix whose elements are proportional to the parameter estimates of heterogeneous parameters from the maximum likelihood estimation on the pooled data (i.e., similar to Manchanda et al. 2004, we ignore the difference between the consumers in this step).

2b. Compute the pseudo-likelihood for consumer i at the candidate vector α_i^{r*}, that is, $q(\gamma^r, \alpha_i^{r*})$, conditional on $DATA_i$ and $E\hat{V}(s, \gamma^r, \alpha_i^{r*})$, which is the Emax function approximated by the weighted average of the N past pseudo-value functions: $E\tilde{V}(s, \gamma^{r*}, \alpha_i^{r*})_{l=r-N}^{r-1}$

$$E\hat{V}(s, \gamma^r, \alpha_i^{r*}) = \sum_{l=r-N}^{r-1} \left[E\tilde{V}(s, \gamma^r, \alpha_i^{r*})\right] \frac{K_h(\gamma^r - \gamma^{l*}, \alpha_i^{r*} - \alpha_i^{l*})}{\sum_{k=r-N}^{r-1} K_h(\gamma^r - \gamma^{k*}, \alpha_i^{r*} - \alpha_i^{k*})}. \tag{C.8}$$

The weights for the N pseudo-value functions are based on the distances between the candidate vector and the previously stored vectors α_i^l. Similar to Step 1b above, we used the Gaussian kernel:

$$K_h(\gamma^r - \gamma^{l*}, \alpha_i^{r*} - \alpha_i^{l*}) = \frac{2\pi}{NPAR} exp\left[\sum_{j=1}^{NPAR_{Homo}} \frac{(\gamma^r - \gamma^{l*})^2}{2h_j} + \sum_{j=1}^{NPAR_{Hetero}} \frac{(\alpha_i^{r*} - \alpha_i^{l*})^2}{2h_j}\right], \tag{C.9}$$

and again we use Silverman's rule of thumb (Silverman 1986) to determine the optimal bandwidth h.

2c. Similarly, we compute the pseudo-likelihood $q(\gamma^r, \alpha_i^{r-1})$ at the previously accepted vector α_i^{r-1}.

Then we determine whether to accept α_i^r based on the following acceptance probability:

$$Prob_{accept} = \min\left[\frac{q(\alpha_i^{r*})\pi(\alpha_i^{r*})}{q(\alpha_i^{r-1})\pi(\alpha_i^{r-1})}, 1\right], \tag{C.10}$$

where $\pi(.)$ is the prior density: $\pi(\alpha_i) \propto (\alpha_i - \Delta^{r-1} Z_i) V_\alpha^{r-1} (\alpha_i - \Delta^{r-1} Z_i)'$.

If accept, set $\alpha_i^r = \alpha_i^{r*}$; if reject, set $\alpha_i^r = \alpha_i^{r-1}$.

Note that Step 2 is iterated for all consumers $i = 1, \ldots, I$, where I is the total number of consumers in the estimation sample.

Step 3. On the basis of α_i^r, draw the posterior mean $\mathbf{\Delta^r}$ from the posterior density:

$$\mathbf{\Delta^r} \sim N(\tilde{\mathbf{\Delta}}, \tilde{V_\alpha}), \tag{C.11}$$

where $\tilde{\Delta} = \tilde{V_\alpha}(A_\alpha^{-1}\bar{\delta} + \Sigma_{i=1}^{I} Z_i' V_\alpha^{r-1} \alpha_i^r)$ and $\tilde{V_\alpha} = (A_\alpha^{-1} + \Sigma_{i=1}^{I} Z_i' \cdot V_\alpha^{r-1} Z_i')$. We set the priors to be uninformative $\bar{\delta} = 0$ and $A_\alpha = \mathbf{100I}$.

Step 4. Draw V_α^r from the inverse Wishart distribution:

$$V_\alpha^r \sim IW\left(\nu + I_{|\alpha|}, \quad \sum_{i=1}^{I} (\alpha_i^r - \Delta^r Z_i)(\alpha_i^r - \Delta^r Z_i)'\right). \tag{C.12}$$

We set $I_{|\alpha|}$, the prior mean of V_α, to $0.1\mathbf{I}$; and the prior degrees of freedom to $\nu = NPAR_{hetero} + 3$, where $NPAR_{hetero}$ is the number of heterogeneous parameters.

Step 5. Use the candidate parameters α_i^{r*} to compute the pseudo-value function $E\tilde{V}(s, \gamma^{*r}, \alpha_i^{*r})$ at the current iteration r. This step uses the Emax approximation computed in Step 2.

Step 6. For each consumer i, compute the pseudo-likelihood at the current iteration r: $q(\gamma^r, \alpha_i^r)$.

Step 7. Use the pseudo-value function and the pseudo-likelihood function from Steps 5 and 6 to update the history $\mathcal{H}^{r+1}$. Go to iteration $r + 1$.

C.6. Estimation Results

C.6.1. Nonparametric First Stage Content Set Transition.

We first estimate Ω using a bin estimator. We use subscripts 1, 2, and 3 to denote the low, medium, and high content states, respectively.[20] Specifically, we find

$$\Omega = \begin{bmatrix} 0.924 & 0.076 & 0 \\ 0 & 0.923 & 0.077 \\ 0 & 0 & 1 \end{bmatrix}.$$

Appendix D. Additional Empirical Analysis

Here we provide some additional data on plan choices and evidence of learning and examine consumer heterogeneity with regard to service time.

D.1. Purchase and Firm Outcomes

We now examine the consumers' purchase decisions, and subsequently firm marketing outcomes (revenue, cost and profit). On average, consumers subscribed to the service for

9.4 months. Low, Medium, and High plans accounted for 14%, 73%, and 13% of purchases, respectively. We examine plan-switching frequencies, summarized in Table D.1. Two patterns emerge. First, plan switching is more likely to occur between adjacent plans but not between the Low and High plans. Second, consumers are more likely to exit from the Low plan, compared with from either the Medium and High plans.[21]

D.1.1. Learning. Past literature has suggested that consumers may learn over time about a service, and multiple sources of learning may exist in other empirical contexts (see Ching et al. 2013 for an excellent review). For example, consumers may learn about the quality (Erdem and Keane 1996, Crawford and Shum 2005, Iyengar et al. 2007, Gopalakrishnan et al. 2014). If there is substantial consumer learning, then we would expect consumers to be more likely to switch plans *earlier* in their tenure. We examine the plan-switching frequencies for each of the first six months of their tenure with the firm. We do not see a clear decreasing trend in the switching or exit probability across these six months. Because the reductions in switching over time are typical variations used to identify the degree of learning, our empirical setting would not be able to identify any learning when no such reduction is present in the data.

D.1.2. Seasonality by Month. We have already seen that at the weekly level, there is a differential level of consumption across weekday versus weekend. We examine the consumption across months to determine whether there are longer-term seasonal consumption patterns that we might need to consider. From Figure D.1, we do not find evidence of significant monthly variation in consumption.

D.1.3. Service Time. Given the previously noted differences in consumption patterns across short versus long service times, we break down the purchase decisions and related firm outcomes by consumer service times. The results are summarized in Table D.2, from which several observations can be made. First, the purchase shares do not vary significantly across these consumers with short and long service times. However, consumers with a shorter service time stay with the firm longer on average (11.6 versus 7.1 months).

D.1.4. Content Heterogeneity. A valid concern is that the switching costs may be conflated with the consumers' heterogeneous viewing preferences. Specifically, the lack of switching may be due to the fact that the consumer is not interested in the newly added content. To examine the possible confounding effect of content heterogeneity, we consider the following regression analyses. In each regression, the dependent variable $DaysKept_{ikt}$ is the number of days movie k was kept by the consumer i at month t [mean (standard deviation) = 9.7 (13.7) days]. The independent variables include dummy variables for movie genres (War is used as the reference genre) and several controls: (1) subscription choice of consumer i in month t; (2) service time of consumer i; and (3) consumer fixed effects (the consumer-specific service time was excluded in this specification). Results are presented in Table D.3. To summarize, none of the coefficients of the genre dummies were found to be significant; thus, genre-dependent utility is unlikely to play a critical role.

Table D.1. Pattern of Plan Switching

Plan	Low	Medium	High	Outside option
Low	88.6	1.3	0.0	10.1
Medium	2.1	86.0	2.4	9.5
High	0.5	3.5	88.4	7.5

Notes. Values are percentages. The (i, j)th entry in each plan-switching matrix is the percentage of plan choices of plan j if the preceding plan is i. The outside option (Exit) is shown in the last column.

We also computed the individual-level Herfindahl indices for consumed movie genres. We find that across the consumers, the indices are not excessively large (mean 0.38, standard deviation 0.23). We treat these as indicative that viewing preference heterogeneity across genres is less likely to be confounded with switching cost in our empirical context. However, modeling genre-specific preferences would be a useful model extension that has not been undertaken in this paper.

D.1.5. Consumption Choices. We expect consumers' consumption decisions to be influenced by the number of available movies from the focal firm (which steadily increased from approximately 300 to approximately 1,300 unique titles during the observation period) and the consumption occasion (i.e., weekdays versus weekends). Observe that inventory represents *current* consumption opportunities, whereas movies in the mail represent *future* consumption opportunities, and a forward-looking consumer accounts for both. To examine whether consumers behave consistently with a forward-looking model, we examine how consumer i's consumption in period (day) t, c_{it} depends on these factors:

$$\begin{aligned} c_{it} = \theta_{0i} &+ \theta_1\ inventory_{it} + \theta_2\ n\_arrivingsoon_{it} \\ &+ \theta_3\ n\_arrivinglate_{it} + \theta_4\ weekend_t \\ &+ \theta_5\ n\_contentset_t + \theta_6\ n\_contentset_t^2 \\ &+ \theta_7\ n\_cumulativeconsumption_{it} + e_{it}. \end{aligned}$$

Consumption (c_{it}) is modeled as a function of the number of movies in the consumer's possession ($inventory_{it}$) and

Figure D.1. Consumption Seasonality by Month

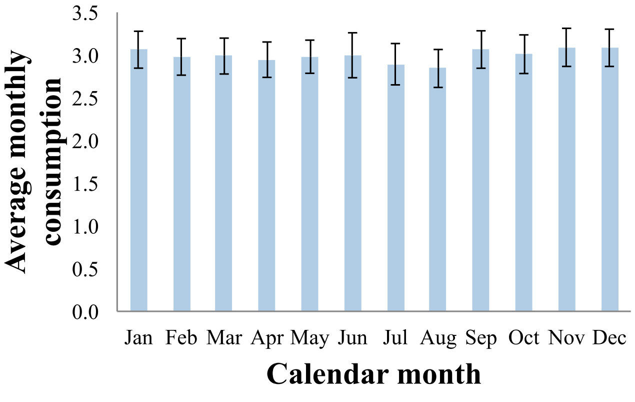


Table D.2. Firm Outcomes by Service Times

	Consumers with Short service				Consumers with Long service			
	Mean	Standard deviation	Min	Max	Mean	Standard deviation	Min	Max
Tenure (months)	11.6	8.3	1	32	7.1	6.7	1	32
Cost ($)	64.8	44.6	2	241.0	29.0	26.7	2	102.7
Revenue ($)	188.7	121.8	19.95	565.9	121.1	105.4	19.95	459.9
Profit ($)	123.9	86.2	10.1	471.4	92.1	82.0	12.0	372.2
	Low plan	Medium plan	High plan		Low plan	Medium plan	High plan	
Purchase shares (%)	14.1	72.9	13.0		14.8	73.8	11.3	

Table D.3. Analyses on Number of Days Kept by the Consumer

	Model 1		Model 2		Model 3	
	Estimate (SE)	p-value	Estimate (SE)	p-value	Estimate (SE)	p-value
Intercept	10.89 (0.972)***	<2e-16	–0.302 (1.251)	0.809		
Action benre	–1.149 (0.999)	0.250	–0.969 (0.992)	0.328	–0.513 (0.904)	0.570
Children genre	0.377 (1.204)	0.754	0.498 (1.194)	0.677	0.051 (1.120)	0.649
Classics genre	–1.159 (2.001)	0.562	–0.812 (1.984)	0.682	0.755 (1.823)	0.678
Comedy genre	–0.955 (1.018)	0.348	–0.710 (1.009)	0.482	–0.274 (0.925)	0.767
Drama genre	–0.439 (1.003)	0.202	–0.226 (0.994)	0.820	0.103 (0.906)	0.909
Romance genre	–1.381 (1.082)	0.199	–1.135 (1.074)	0.291	–0.439 (0.982)	0.655
Sci-fi genre	–1.452 (1.131)	0.199	–1.425 (1.122)	0.204	–0.610 (1.025)	0.552
Suspense genre	–1.963 (1.061)	0.064	–1.682 (1.053)	0.110	–0.324 (0.962)	0.736
Plan choice & service time	Not included		Included		Not included	
Individual FE	Not included		Not included		Included	
Adjusted R^2	0.0007		0.017		0.476	

Note. FE, fixed effect; SE, standard error.
***$p < 0.001$.

the number of movies that she will receive the next day ($n\_arrivingsoon_{it}$) or after the full service time ($n\_arrivinglate_{it}$). Forward-looking consumers will account for future availability of movies and adjust consumption decisions (e.g., a consumer with movies arriving soon is more likely to consume, compared with the case in which she has movies arriving later). The dummy variable $weekend_t$ is 1 if day t is either Saturday or Sunday, and 0 otherwise; it captures the systematic difference in consumption utility on weekends versus weekdays. The $n\_contentset_t$ denotes the number of unique movie titles available on day t. This variable and its squared term account for the potential effects of a larger content set on the average consumption.

Results are summarized in Table D.4. We view these results as factors supporting our decisions in the structural model. *However, we do not view the estimates causally, because the numbers of movies arriving at a specific period are endogenous variables.* We find that movie inventory has a significant positive effect on consumers' consumption decision. Consistent with forward-looking consumers, movies in the mailing process also have significant effects (e.g., a consumer is more likely to consume when she expects more movies to be arriving soon). Consistent with model-free evidence, consumers are more likely to consume during weekends. The content set size also has a positive effect on the consumption.

Cumulative consumption has a positive and significant impact, but the magnitude is quite small; that is, with an additional 25 movies watched (the total number of movies watched by an average consumer in her entire lifetime in our data period), the consumption probability changes from 9.8% to 10.3%.

Table D.4. Daily Consumption Regression

Variable	Coefficient	t-value
Number in consumer's inventory	4.39×10^{-2}***	31.49
Number arriving soon	3.00×10^{-2}***	24.68
Number arriving late	-1.69×10^{-2}***	–5.99
Weekend dummy	4.31×10^{-2}***	18.61
Content set size	2.94×10^{-4}***	3.40
Square of content set size	-2.35×10^{-7}**	–2.86
Cumulative number of movies watched	2.34×10^{-4}*	2.35

Notes. Dependent variable: number of movies c_{it} watched on day t. Individual consumer fixed effects included.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Endnotes

[1] See https://www.cnbc.com/2018/05/22/rent-the-runway-2018-disruptor-50.html and http://www.scw-mag.com/sections/retail/868-rent-the-runway.

[2] A consumer can sign onto her online account to add to her personal movie queue (i.e., the set of movies that she would like the firm to send to her). Movies that have been sent to the consumer are removed from the queue.

[3] We also examine whether consumers demonstrate learning and how service time impacts consumers' choices and firm outcomes. These are detailed in Appendix D.

[4] We also estimate a semiparametric model with the following utility specification: $u(c_{it}) = \theta_1 \mathbf{I}[c_{it} = 1] + \theta_2 \mathbf{I}[c_{it} = 2] + \theta_3 \mathbf{I}[c_{it} = 3] + \alpha_3 c_{it} v_{it} + \alpha_w c_{it}\ \mathbf{I}[t \in \mathbf{T_w}] + \alpha_{cs} c_{it}\ \log(\omega_t)$, with the rest of the utility terms the same as the current model.

[5] The linear-quadratic form can be viewed as an approximation to several more complex utility functions.

[6] We add a *dormant state* that allows consumers to choose the outside option and remain in our tracking. This state is an alternative to permanent exit and allows the consumer to return by subscribing to a plan in the future.

[7] We estimate $\boldsymbol{\Omega}$ separately from the data and incorporate it into consumers' expectations.

[8] Overall, we have one endogenous state (x_{it}) and three exogenous states, and all except ω_t evolve deterministically. The total state space including both short-run and long-run contains (1) the mailing state, which characterizes the consumers' DVD inventory and the number of DVDs at different stages of the mailing process; (2) the day of week, which affects immediate consumption utility; (3) the day in the payment cycle; and (4) the size of the content set.

[9] Overall, there are 247 (N_x) ×7 (number of days in the week) ×30 (number of days in a payment cycle) ×3 (N_ω) = 155,610 short-run states. We show a representative value function in Appendix C.

[10] From a computational perspective, it is useful to separate out the long-run and short-run value functions. Consumption choice data are used to identify the short-run consumption preference parameters, whereas the combination of the short-run value function, content set changes, and plan choices serves to identify the long-run parameters. Moreover, in computing the counterfactuals, this is especially helpful because the price only impacts the long-run value function and not the short-run value; we only need to compute the compute the short-run value function once (for each service time) rather than for every possible price vector. We thank an anonymous reviewer for this suggestion.

[11] The estimation procedure was written in *Julia* for computational speed; a single iteration of the IJC algorithm requires approximately 10 seconds on a 64-core Amazon cloud computer. We ran 30, 000 iterations for all models and computed parameter estimates based on the last 10,000 iterations. Figure C.2 in Appendix C plots the pseudo-likelihood across iterations and shows that model estimates converged reasonably quickly.

[12] The focal firm provided us with this cost estimate, where \$0.9 is spent the two-way mailing cost and \$1.1 spent on handling of each DVD.

[13] In practice, the firm might need to purchase more copies of movies, especially popular ones. Our estimate of costs might likely prove to be an underestimate, and thus a reduction in service time might prove even worse from a profitability viewpoint.

[14] To see a simple example of this, consider a monopolist facing two equal-sized (normalized to 1) groups of consumers with valuation $\mu - \delta$ and $\mu + \delta$, where δ characterizes the variation in valuations. The firm faces zero costs. When variation is low (i.e., $\delta < \frac{\mu}{3}$), the firm's revenue (or profit) $\pi^* = 2(\mu - \delta)$ is decreasing in δ. Further, observe that it is easy to construct cases in which, with an increase in both μ and δ, the revenue decreases.

[15] See http://www.nbcnews.com/id/11262292/ns/business-us_business/t/frequent-netflix-renters-sent-back-line/.

[16] As of 2018, Netflix streaming now accounts for approximately one-third of the entire U.S. online traffic. Source: http://www.streamingmedia.com/Articles/Editorial/Featured-Articles/Stream-This!-Netflixs-Streaming-Costs-65503.aspx. Retrieved August 10, 2018.

[17] The streaming cost for a two-hour movie is estimated to be six cents for the standard format movie, and nine cents for the high-definition movie. The licensing fee typically varies across movies and can be as high as \$4 per movie. Source: http://www.streamingmedia.com/Articles/Editorial/Featured-Articles/Stream-This!-Netflixs-Streaming-Costs-65503.aspx. Retrieved August 10, 2018.

[18] At each level of marginal cost, we conduct a grid search over 1,000 price points uniformly distributed between \$1 and \$100 for the subscription prices and between \$0 and \$20 for the per-movie prices, respectively.

[19] For example, 100,000 unique titles are available from Netflix, compared with a few thousand in a retail store.

[20] Note that $\Omega_{33} = 1$ is fixed, not estimated.

[21] We also separately examined the switching matrices for periods with a high content set size (more than 650 movie titles) and a low content set size (fewer than 650 titles) and found the same patterns. Furthermore, the probability of exiting is significantly lower when the content set is high (versus low).

References

Allenby G, Rossi PE (1998) Marketing models of consumer heterogeneity. *J. Econometrics* 89(1):57–78.

Andrews RL, Ansari A, Currim IS (2002) Hierarchical Bayes versus finite mixture conjoint analysis models: A comparison of fit, prediction, and partworth recovery. *J. Marketing Res.* 39(1):87–98.

Armstrong M, Rochet J-C (1999) Multi-dimensional screening: A user's guide. *Eur. Econom. Rev.* 43(4-6):959–979.

Bassamboo A, Kumar S, Randhawa RS (2009) Dynamics of new product introduction in closed rental systems. *Oper. Res.* 57(6): 1347–1359.

Cachon GP, Feldman P (2011) Pricing services subject to congestion: Charge per-use fees or sell subscriptions? *Manufacturing Service Oper. Management* 13(2):244–260.

Ching AT, Erdem T, Keane MP (2013) Invited paper-learning models: An assessment of progress, challenges, and new developments. *Marketing Sci.* 32(6):913–938.

Crawford GS, Shum M (2005) Uncertainty and learning in pharmaceutical demand. *Econometrica* 73(4):1137–1173.

Crawford GS, Shum M (2007) Monopoly quality degradation and regulation in cable television. *J. Law Econom.* 50(1):181–219.

Erdem T, Keane MP (1996) Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing Sci.* 15(1):1–20.

Goettler RL, Clay K (1997) Tariff choice with consumer learning and switching costs. *Econometrica J. Econometric Soc.* 65(3):487–516.

Gopalakrishnan A, Iyengar R, Meyer RJ (2014) Consumer dynamic usage allocation and learning under multipart tariffs. *Marketing Sci.* 34(1):116–133.

Handel BR (2013) Adverse selection and inertia in health insurance markets: When nudging hurts. *Amer. Econom. Rev.* 103(7):2643–2682.

Imai S, Jain N, Ching A (2009) Bayesian estimation of dynamic discrete choice models. *Econometrica* 77(6):1865–1899.

Iyengar R, Ansari A, Gupta S (2007) A model of consumer learning for service quality and usage. *J. Marketing Res.* 44(4):529–544.

Lambrecht A, Seim K, Skiera B (2007) Does uncertainty matter? Consumer behavior under three-part tariffs. *Marketing Sci.* 26(5): 698–710.

Magnac T, Thesmar D (2002) Identifying dynamic discrete decision processes. *Econometrica* 70(2):801–816.

Manchanda P, Rossi PE, Chintagunta PK (2004) Response modeling with nonrandom marketing-mix variables. *J. Marketing Res.* 41(4):467–478.

Maskin E, Riley J (1984) Monopoly with incomplete information. *RAND J. Econom.* 15(2):171–196.

McManus B (2007) Nonlinear pricing in an oligopoly market: The case of specialty coffee. *RAND J. Econom.* 38(2):512–532.

Milkman KL, Rogers T, Bazerman MH (2009) Highbrow films gather dust: Time-inconsistent preferences and online DVD rentals. *Management Sci.* 55(6):1047–1059.

Miller NH, Osborne M (2014) Spatial differentiation and price discrimination in the cement industry: Evidence from a structural model. *RAND J. Econom.* 45(2):221–247.

Mussa M, Rosen S (1978) Monopoly and product quality. *J. Econom. Theory* 18(2):301–317.

Netzer O, Lattin JM, Srinivasan V (2008) A hidden Markov model of customer relationship dynamics. *Marketing Sci.* 27(2):185–204.

Ngwe D (2017) Why outlet stores exist: Averting cannibalization in product line extensions. *Marketing Sci.* 36(4):523–541.

Randhawa RS, Kumar S (2008) Usage restriction and subscription services: Operational benefits with rational users. *Manufacturing Service Oper. Management* 10(3):429–447.

Rochet J-C, Stole LA (2002) Nonlinear pricing with random participation. *Rev. Econom. Stud.* 69(1):277–311.

Rossi PE, Allenby GM, McCulloch R (2012) *Bayesian Statistics and Marketing* (John Wiley & Sons, Hoboken, NJ).

Shiller B (2014) First degree price discrimination using big data. Technical report, Brandeis University, Waltham, MA.

Shum M (2004) Does advertising overcome brand loyalty? Evidence from the breakfast-cereals market. *J. Econom. Management Strategy* 13(2):241–272.

Silverman BW (1986) *Density Estimation for Statistics and Data Analysis*, vol. 26 (CRC Press, Boca Raton, FL).

Stole L (2003) Price discrimination and imperfect competition. Armstrong M, Porter R, eds. *Handbook of Industrial Organization*, vol. 3 (Elsevier, Amsterdam), 34–47.

Tong C, Rajagopalan S (2014) Pricing and operational performance in discretionary services. *Production Oper. Management* 23(4): 689–703.