Online Appendix

A Exogenous Search

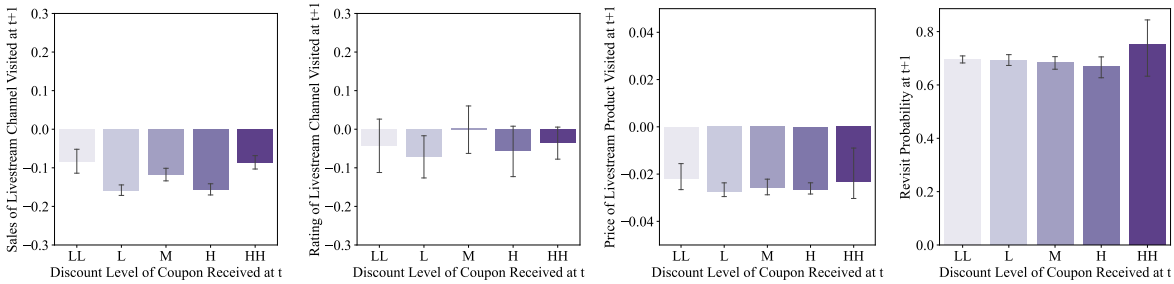
Our current setting assumes that consumers' search behaviors are independent of the platform's coupon allocations. In this section, we provide empirical evidence of this assumption. Fig.A1 plots the relationship between the discount level of the coupon received by a consumer in purchase incidence t and her subsequent search behavior in incidence $t+1$.[42] Specifically, we consider four search behaviors that could plausibly be affected by the previous coupon allocation: the seller's sales, the seller's rating, the average price of the products in the channel, and the probability that the consumer has visited the channel before (i.e., the revisit probability). For each behavior, we conduct an analysis of variance (ANOVA).

In panel (a), we find no significant difference in the sales of the seller visited in incidence t+1 based on the discount level of the coupon received in incidence t (F-stat = 1.491, p_value = 0.202). In other words, a consumer who receives a big discount now is not more or less likely to seek out a channel with high or low sales in the next search. We find analogous results in the remaining panels. In panel (b), we find no significant difference in the average rating of the seller visited in incidence t+1 based on the discount level of the coupon received in incidence t (F-stat = 1.643, p_value = 0.160). In panel (c), we find no significant difference in the average price of the product considered in incidence t+1 based on the discount level of the coupon received in incidence t (F-stat = 2.016, p_value = 0.089). In other words, there is no evidence that consumers switch to highly (or poorly) rated sellers or to low-priced (or high-priced) products after receiving a deep discount. Finally, we test whether consumers are more likely to revisit a particular host after receiving a high-discount coupon from this host. In panel (d), we find no significant difference in the likelihood that the consumer is revisiting a channel in incidence t+1 based on the discount level of the coupon received in incidence t (F-stat = 1.369, p_value = 0.242). The revisit probability is always around 70%.

[42] We fix the coupon's threshold level at M to ensure a fair comparison. This rule applies to all the subplots in Fig.A1.

Figure A1: Evidence of Consumer Exogenous Search Behavior

(a) Seller Sales_t+1 vs. Coupon Discount Level_t (b) Seller Rating_t+1 vs. Coupon Discount Level_t (c) Product Price_t+1 vs. Coupon Discount Level_t (d) Revisit_t+1 vs. Coupon Discount Level_t



We acknowledge that consumers' search behaviors could be influenced by coupon allocations in other marketing contexts, in which case the researchers could re-categorize the state variables associated with the host, product, and livestream to be dynamic instead of static.

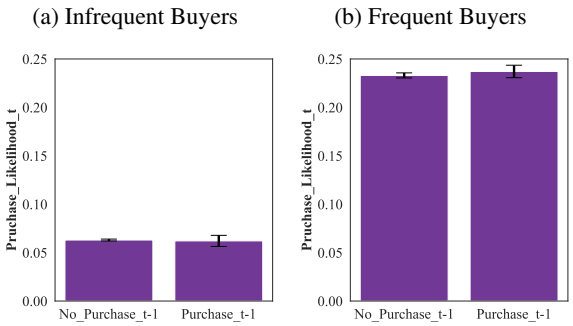
B Disentangling Unobserved Heterogeneity and State Dependence

We address a concern about the test for state dependence: if the conditional purchase probability is higher than the marginal purchase probability, it may reflect not structural state dependence or inertia but rather the confounds of unobserved heterogeneity. That is, consumers may differ along a serially correlated, unobserved propensity to make purchase decisions (Heckman 2007). We use the method proposed in Dubé et al. (2010) to tease apart state dependence from unobserved heterogeneity. Specifically, we rely on spells during which the consumer first made a purchase initiated by a discount (defined as a purchase for which the consumer redeemed a coupon with the highest discount level, HH) and then continued to purchase after prices returned to "typical" levels (a coupon with discount level L or lower). We compare this repeat-purchase rate with the marginal purchase probability, and we find no significant difference (10.86% vs. 10.79%, p_value = 0.65). We conclude that state dependence is absent even after controlling for unobserved heterogeneity.

C Additional Evidence of the Absence of Variety-Seeking

Although we did not find evidence of variety-seeking in our initial analysis, it is possible that variety-seeking behavior occurs only among heavy repurchasers. To test this hypothesis, we recreate the variety-seeking plot with two groups of consumers: infrequent buyers (< 5 purchases in the sample period) and frequent buyers (at least 5 purchases in the sample period). Fig.A2 shows that the purchase likelihood is not negatively influenced by prior purchase experience for either group of consumers, indicating an absence of variety-seeking behavior.

Figure A2: Coupon Redemption Rate by Prior Purchase Experience and Purchase Frequency



D Forward-Looking Behavior

D.1 Evidence of the Absence of Consumer Forward-Looking Behaviors

To assess whether consumers in our sample are forward-looking, we compare the purchase frequency under three scenarios: (1) the consumer receives coupons with a stable discount level over time, (2) the consumer receives coupons with an increasing discount level over time, and (3) the consumer receives coupons with a decreasing discount level over time. If a forward-looking consumer anticipates a deeper discount (i.e., price reduction) in the future, she may strategically wait to make a purchase, leading to a lower initial purchase frequency than a consumer who receives the same discount every time (Su 2007). On the other hand, if a forward-looking consumer anticipates a smaller discount (i.e., price surge) in the future, she may accelerate purchases now to stockpile (Neslin et al. 1985).

We formally test for strategic waiting and stockpiling behaviors in Table A1. Because the purchase frequency is heavily influenced by the discount level, as demonstrated in Fig.2a, we conduct separate analyses at three of the five discount levels (based on the first coupon received): L, M, and H. The purchase time (mean and standard deviation) for the three scenarios are displayed in Table A1. For example, when the starting discount level is L and the discounts are increasing (i.e., scenario 2), the average purchase frequency in the beginning (when consumers receive the initial, low-discount coupon) is 0.047.[43] When consumers receive stable, low-level coupons (i.e., scenario 1), the average initial purchase frequency is 0.043, which is not significantly different from the purchase frequency in scenario 2 (t-test statistic = 0.389, p-value = 0.697). Therefore, we find no evidence of strategic waiting. Similarly, when the starting discount level is L and discounts are decreasing (i.e., scenario 3), the average initial purchase frequency is 0.033, which also is not significantly different from that in scenario 1 (p_value = 0.294), suggesting an absence of stockpiling. We find the same null results with the starting discount levels of M and H.

[43] A purchase frequency of 0.043 means 4.3 purchases for every 100 incidences.

Table A1: Evidence of No Forward-Looking Behaviors

Starting Discount Level	Purchase Frequency at the Starting Discount Level	Discount Level Trend Over Time		
		(1) Stable	(2) Increasing	(3) Decreasing
L	Mean	0.043	0.047	0.033
	Standard Deviation	0.079	0.119	0.088
	T-test (vs. Stable)		0.389	-1.050
	P-value (vs. Stable)		0.697	0.294
H	Mean	0.111	0.112	0.093
	Standard Deviation	0.120	0.286	0.255
	T-test (vs. Stable)		0.033	-0.509
	P-value (vs. Stable)		0.974	0.612

The results in Table A1 suggest that consumers in our context are myopic rather than forward-looking; they do not engage in strategic waiting or stockpiling based on assumptions about future prices. The absence of forward-looking behaviors may be attributable to characteristics of the livestream shopping platform, where the best-selling product categories (apparel, cosmetics, and jewelry) are highly seasonal, so consumers might not want to delay purchases (and miss the fashion trend) or stockpile products (which soon would be obsolete).

Although the consumers in our data do not seem to be forward-looking, we acknowledge the general importance of incorporating consumer strategic behaviors into the designs of dynamic coupon targeting policies. Indeed, in many marketing contexts, it is easier for consumers to learn and form expectations about price or quality. Next, we discuss how to extend our reinforcement learning model to such contexts.

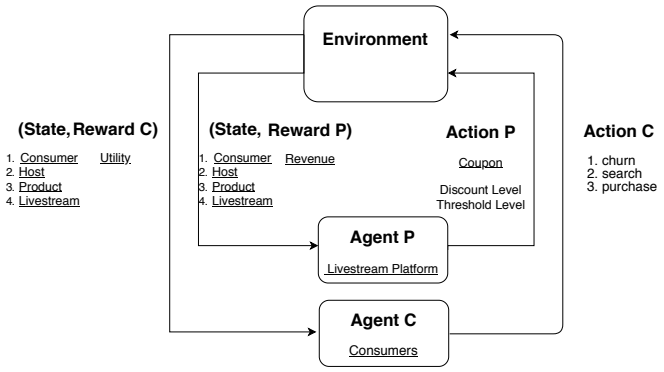
D.2 Multi-Agent Reinforcement Learning

To accommodate consumer forward-looking behaviors, future research could explore multi-agent reinforcement learning (Littman 1994), which allows consumers and the platform to be independent agents. Each agent maximizes its own total discounted rewards, and the reward function of each agent depends on the actions of all agents. In this section, we briefly define multi-agent reinforcement learning and explain how to use it in our context. We leave the solution and estimation for future research.

We use Fig.A3 to illustrate how a Markov game can solve the dynamic targeting problem. There are two agents, the platform P and consumers C. At time t, when a consumer enters a livestream channel, the platform agent takes an action: it chooses a coupon for the consumer in this purchase incidence. After observing the platform's action, the consumer also takes an action: she decides whether to purchase and whether to churn after this incidence. The joint actions of the platform and the consumer determine the reward for each (revenue for the platform; utility from the purchase incidence for the consumer). Then, the environment evolves to time $t+1$, and both the platform and consumer observe the next state, which

includes features of the consumer, host, product, and livestream.

Figure A3: Multi-Agent Reinforcement Learning Framework



The Markov game illustrated in Fig.A3 allows both the platform and consumers to learn from experience collected through trial-and-error strategies. Specifically, if the platform's policy is to increase the discount level (i.e., decrease prices) over time, then consumers will incorporate this policy into their purchase decisions and engage in strategic waiting because the platform's policy is part of the value function for consumers.

D.3 Optimal Pricing Strategy for Forward-looking Consumers

What would be the optimal pricing strategy for forward-looking consumers? We draw on the theoretical literature on dynamic pricing to offer some conjectures. The optimal strategy depends on the primary intertemporal tradeoff. If the reference price effect is the primary intertemporal tradeoff, then Wu et al. (2015) show that price skimming (i.e., markdown price) is the optimal pricing strategy with forward-looking consumers who form price expectations or reference prices based on the retailer's actions in the past. The presence of strategic consumers makes the firm discount less over time, leading to a more stable markdown price. If loyalty or inertia is the primary intertemporal tradeoff, then Klemperer (1987) shows that price penetration is the equilibrium optimal strategy under the assumption that consumers are forward-looking and have rational expectations for future prices. The reason is that firms will raise their prices in the second period to take advantage of the fact that first-period customers are now locked in. Firms may charge higher prices in the first period with forward-looking consumers than with myopic consumers because the demand of forward-looking consumers is less elastic; they recognize that a firm with a lower first-period price will gain a greater market share and will charge a higher price in the second period, so forward-looking consumers are wary of becoming attached to the supplier. Finally, Seetharaman and Che (2009) show that if variety-seeking is the primary intertemporal tradeoff, then collusive, enduring high prices are optimal even for forward-looking consumers. In fact, first-period prices increase even more for rational, forward-

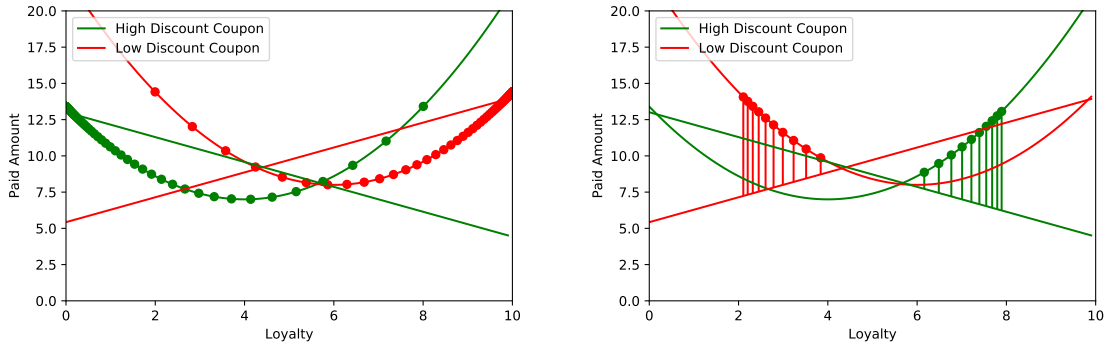
looking consumers than for myopic consumers because forward-looking consumers recognize that they will be partially locked in to an untried supplier in the second period, so they must predict second-period prices when making their first-period purchase decisions, and the prediction makes the first-period price cut less attractive.

E An Example of Model Bias

This section uses an illustrative example (Fig.A4) to explain the intuition behind model bias, and we present empirical evidence of potential model bias.

Figure A4: Illustration of Model Bias

(a) Observed Policy (b) New Policy



Imagine there are two types of coupons, one with a high discount level, color-coded in green, and the other with a low discount level, color-coded in red. Fig.A4 exhibits the relationship between loyalty and the paid amount (revenue) for purchases made with the two types of coupons. We define "loyalty" as the consumer's number of past purchases (e.g., a loyalty of 2 means that the consumer has bought twice before). It is possible that, under a certain behavioral mechanism, we could observe a non-linear, quadratic relationship between loyalty and the paid amount (shown as the green curved line and the red curved line). Specifically, when loyalty is low, revenue decreases with purchase experience; when loyalty is high, revenue increases with purchase experience. Also, it is possible that the high-discount coupon generates higher revenue from high-loyalty customers, whereas the low-discount coupon generates higher revenue from low-loyalty customers. Thus, in Fig.A4, the green curved line is above the red curved line when loyalty is higher than 6.

We further assume that in the observed data (panel (a)), the firm's strategy rewards new consumers by giving them more high-discount (green) coupons, whereas loyal consumers receive more low-discount (red) coupons. In the plot, each dot is one observation; the green dots (high discounts) are concentrated in the low-loyalty region, and the red dots (low discounts) are concentrated in the high-loyalty region. Thus,

although the true relationship between loyalty and revenue in the data generating process is quadratic, an econometrician might instead assume that a linear relationship (the straight lines) is the best fit. The linear relationships are estimated using ordinary least squares (OLS) and are unbiased in the current data, under the current policy.

Next, we consider a new policy (panel (b)) in which the firm reverses its previous strategy and now gives more high-discount coupons to loyal customers instead of to new customers (the green dots cluster in the high loyalty region in panel (b)). Now, the revenue predicted by the linear model is negatively biased because the green straight line lies below the green curved line. The biased predictions lead to suboptimal strategy decisions. For example, in the high-loyalty region (loyalty > 6), the linear model predicts that the low-discount coupon is the best strategy (the straight red line is above the straight green line), while the quadratic model predicts the opposite (the curved green line is above the curved red line).

The example demonstrates two sources of model bias. The first source of bias is distribution mismatch or covariate shift. The reward predictor (straight lines) was trained on the past data and can perform well (i.e., be unbiased) on past data but not necessarily on data generated using a new policy. The reward predictor is formed without knowledge of the new policy and, hence, when approximating the reward, the predictor might focus too much on areas that are irrelevant for the new policy and not enough on areas that are important for the new policy (Dudík et al. 2014). In our setting, although the platform randomly distributed coupons, the probabilities associated with different coupon types are not uniform. As shown in Table 2, there are more coupons with the lowest discount level (0-15%) than with the highest discount level (85%-100%) because the platform cannot afford to offer large markdowns as often as small markdowns. So, a model built on our observed data could suffer from the first source of model bias. The second source of model bias is wrong functional forms. Consumer behaviors in the information-rich environment are challenging to model. The straight line here is an extreme example, but even flexible non-parametric models might not accurately capture the intricacies in the true data-generating process. In fact, many prior studies on consumer click-through rates and purchase likelihoods present predictions with only single-digit precision (Wen et al. 2019).

F CCP

Although both Q-learning and the Conditional Choice Probability (CCP) estimator (Hotz and Miller 1993) use sample observations (the so-called cell estimators) in the estimation procedure, they are fundamentally

different algorithms. We compare the properties of the two algorithms on three dimensions: (1) reinforcement learning primitives, (2) how the sample observations are used, and (3) the researcher's role.

First, CCP and Q-learning belong to two different subcategories of reinforcement learning and, therefore, have different primitives. CCP is an estimation algorithm for the exact dynamic programming problem, which assumes that the agent knows the environment, that is, the state transition process and the reward function. Therefore, the agent can solve the dynamic programming program and derive the optimal policy. The data generating process assumes that the agent takes actions according to the optimal policy in each step. By contrast, Q-learning is a solution to the model-free reinforcement learning problem, which assumes that the agent does not know how the environment operates, so the agent learns about the environment while gaining experience. In each step, the agent takes actions according to the policy the agent has learned, which might not yet be optimal. Moreover, CCP assumes that the agent's reward is defined by a utility function that is not observed directly in the data, while the reward for Q-learning is observed in the data.

Second, both CCP and Q-learning use sample observations, but in different fashions. In CCP, sample observations are used to calculate the conditional choice probabilities (hence, the name of the method): the action probabilities in each state and the state transition probabilities for each state-action pair. Then, the CCPs are plugged into the likelihood function in the maximum likelihood estimation routine. On the contrary, in Q-learning, sample observations are used to get the values of the reward and next state, which are plugged into the value iteration update function to iteratively calculate the Q-function (step 9 in Algorithm 1).

Third, the researcher plays different roles when applying CCP and Q-learning. In CCP, the researcher is the econometrician, whose objective is to infer or estimate the parameters in the agent's utility function. When using Q-learning, however, the researcher is the agent herself. She directly observes the reward and creates the optimal policy by learning the value function iteratively. Because Q-learning belongs to the family of model-free reinforcement learning, no model or model parameter is involved. In the tabular case, when the state space is small, no parameter needs to be estimated at all. When the state space is large, the researcher needs to estimate the parameters in the functional approximator, such as DNNs.

To sum up, although both CCP and Q-learning use cell estimators, they differ drastically on the three dimensions mentioned above. For a more detailed comparison of exact dynamic programming and model-free reinforcement learning, see Sutton and Barto (2018).

G Full Structural Model

This section presents a full structural model for consumers' sequential search, purchase, and churn behaviors. An alternative model without search is introduced in §L.

G.1 Model

Assume that all livestream channels are displayed in order on the list/search page, which shows some information about each livestream channel, including the thumbnail, topic, seller's name, average price, number of viewers, and number of likes. When a consumer clicks a channel's link, she enters the channel and can obtain more information such as the livestream content, detailed product descriptions, and a coupon. On a single day, consumers can visit multiple channels that belong to multiple product categories, and they can make multiple purchases. We treat channel visits in different categories as separate, independent search sessions. Formally, a search session is confined to one product category and, at most, one purchase, but it can include multiple clicks (channel visits) in the same category. A search session can end in one of two ways: the consumer makes a purchase, or the consumer switches from one product category to another (which both terminates the first search session and starts a new one). After examining the data, we found that, within one session, consumers rarely revisited a channel that they searched earlier in the session; all purchases were made in the last channel visited. Therefore, in each step of the decision process, the consumer's decisions are (1) whether to stop searching (churn), (2) if not churning, then which channel to search (search), and (3) whether to buy a product from the searched channel (purchase). Our setting can be characterized by a sequential search model without recall (McCall 1970).

G.1.1 Timing

A search session g can consist of multiple coupon reception incidences t. During a search session g, when a consumer visits channel j and receives a coupon, this constitutes a coupon reception incidence t. This timing definition is consistent with that in §4.

G.1.2 Search and Purchase Model

Let w_{ijg} be consumer i's valuation of the list-page information for channel $j \in \{1,\ldots,J\}$ in search session g, and v_{ijg} is the valuation after clicking and watching the livestream channel. Then, w_{ijg} is the expected utility before search, and v_{ijg} is the expected utility from search. w_{ijg} depends on $\mathbf{x}_{ijg}$, a vector of state variables related to the channel (seller) and product (e.g., average price, number of viewers, number of likes). We

assume v_{ijg} follows a normal distribution with a mean of zero and standard deviation σ_j.

Let c_{ijg} be the search cost, which includes the time and mental cost of processing information during search. The search cost depends on $\mathbf{z}_{ijg}$, a vector of consumer- and seller/product-related state variables, such as consumer demographics and the ranking position of the channel. If a consumer chooses not to make a purchase on the platform, she can take the outside option $j = 0$ with a known expected utility u_{it0} (i.e., it does not require search). Then, after search, the realized utility of channel j becomes

$$\begin{aligned} u_{ijg} &= w_{ijg} + v_{ijg}, v_{ijg} \sim N\left(0, \sigma_j^2\right) \\ w_{ijg} &= \mathbf{x}_{ijg}^{'} \boldsymbol{\beta}_i \\ c_{ijg} &= \mathbf{z}_{\mathbf{ijg}}^{'} \boldsymbol{\gamma}_i \end{aligned} \tag{10}$$

The variables ($\mathbf{x}_{ijg}$) that affect the search utility u_{ijs} include those observed after clicking because we assume that $\mathbf{x}_{ijg}$ can be decomposed into two subsets, $\mathbf{x}_{ijg}^s$ and $\mathbf{x}_{ijg}^p$, where $\mathbf{x}_{ijg}^s$ are the variables observed in the search stage (before clicking; the superscript s denotes the search stage) and $\mathbf{x}_{ijg}^p$ are the variables observed in the purchase stage (after clicking; the superscript p denotes the purchase stage). Because $\mathbf{x}_{ijg}^p$ is not observed in the search stage, we can assume that consumers form an expectation about $\mathbf{x}_{ijg}^p$, denoted $\tilde{\mathbf{x}}_{ijg}^p$; thus, in the search stage, $u_{ijg} = \mathbf{x}_{ijg}^s \boldsymbol{\beta}^s + \tilde{\mathbf{x}}_{ijg}^p \boldsymbol{\beta}^p + v_{ijg}$. If we assume that consumers have rational expectations (i.e., the expected variable values deviate from the observed variable values only by a random noise term, that is, $\tilde{\mathbf{x}}_{ijg}^p = \mathbf{x}_{ijg}^p + \boldsymbol{\iota}_{ijg}$ and $\boldsymbol{\iota}_{ijg} \sim N(\mathbf{0}, \Sigma_\iota)$), then the utility function becomes $u_{ijg} = \mathbf{x}_{ijg}^s \boldsymbol{\beta}_i^s + \mathbf{x}_{ijg}^p \boldsymbol{\beta}_i^p + \iota_{ijg}^{'} \boldsymbol{\beta}^p + v_{ijg}$. Let $\varepsilon_{ijg} = \iota_{ijg}^{'} \boldsymbol{\beta}_i^p + v_{ijg}$, then $u_{ijg} = \mathbf{x}_{ijg}^s \boldsymbol{\beta}_i^s + \mathbf{x}_{ijg}^p \boldsymbol{\beta}_i^p + \varepsilon_{ijg} = \mathbf{x}_{ijg}^{\prime} \boldsymbol{\beta}_i + \varepsilon_{ijg}$, $\varepsilon_{ijg} \sim N\left(0, \beta^p \Sigma_\iota \boldsymbol{\beta}^p + \sigma_j^2\right)$. The utility function implies that the variables in equation 10 can include all the covariates, regardless of whether they are observable before clicking. Importantly, $\mathbf{x}_{ijg}$ includes $\mathbf{A}_{it}$, the coupon the consumer receives.[44] For ease of explanation, we can assume that $\mathbf{x}_{ijg} = \begin{pmatrix} \mathbf{S}_{it} \\ \mathbf{A}_{it} \end{pmatrix}$, where $\mathbf{S}_{it}$ is the vector of the state variables and $\mathbf{A}_{it}$ is the vector of action dummy variables. Note that the state variables $\mathbf{S}_{it}$ include those that can incorporate consumer intertemporal tradeoffs (Table A10), such as the reference price effect and state dependence, so the structural model is consistent with the model free evidence documented in §3.4.

We allow for consumer heterogeneity using a random coefficient approach where both the individual utility parameters $\boldsymbol{\beta}_i$ and cost parameters $\boldsymbol{\gamma}_i$ follow a normal distribution; that is, $\boldsymbol{\beta}_i \sim N\left(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta\right)$ and $\boldsymbol{\gamma}_i \sim N\left(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_\gamma\right)$, where $\boldsymbol{\mu}_\beta$ and $\boldsymbol{\mu}_\gamma$ capture the mean effect across consumers, and $\boldsymbol{\Sigma}_\beta$ and $\boldsymbol{\Sigma}_\gamma$ (assumed to be

[44] This assumption contradicts the exogenous search evidence in Appendix A. Future research could separate the utility function into the search utility and purchase utility (Seiler 2013).

diagonal) capture the variance.

The consumer's optimal search strategy can be summarized by the following rules:

1. Selection Rule: If a channel is to be searched, it should be the channel with the highest reservation utility R_{ij}. The reservation utility is defined as the utility that makes the consumer indifferent between choosing the last-searched option (with utility R_{ij}) and continuing with the next search. The reservation utility solves the equation $\int_{R_{ijg}}^{\infty}(u_{ijg}-R_{ijg})\,dF(u_{ijg})=c_{ijg}$, where the left-hand side is the marginal utility of searching channel j, and the right-hand side is the marginal cost of search.

2. Stopping Rule: Terminate search whenever the current realized utility exceeds the reservation utility of every unsearched channel.

3. Purchase Rule: Once the consumer stops searching, compare the last-searched channel with the outside option. If the realized utility of the last-searched channel is higher than that of the outside option, then purchase. Otherwise, take the outside option.

G.1.3 Churn Model

The churn model is specified as a binary logit model. Consumer i's utility of churn from the platform in incidence t is ϕ_{ijg1}, and the utility of staying is ϕ_{ijg0}

$$\begin{aligned}\phi_{ijg1} &= \mathbf{S}_{ijg}\rho+\varepsilon_{ijg1}, \varepsilon_{ijg1}\sim Gumbel(0,1)\\ \phi_{ijg0} &= \varepsilon_{ijg0}, \varepsilon_{ijg0}\sim Gumbel(0,1)\end{aligned}$$

G.2 Estimation

The likelihood function for the structural parameters $\boldsymbol{\theta}=\left(\boldsymbol{\mu}_{\beta},\boldsymbol{\Sigma}_{\beta},\boldsymbol{\mu}_{\gamma},\boldsymbol{\Sigma}_{\gamma},\left\{\sigma_j\right\}_{j=1}^{J},\rho\right)$ can be specified based on the search rules. The selection rule implies that, in the j_{th} search, the reservation utility for the channel is higher than that of all the to-be-searched and unsearched channels. Thus, the corresponding likelihood function is $L_{ijg}^{Selection}=Pr\left(R_{ijg}\geq R_{ikg},\forall k\in\{j+1,\ldots,J\}\right)$. The stopping rule implies that 1) the reservation utility for the channel in the j_{th} search must be higher than the realized utilities of all previously searched channels; otherwise, the consumer would have stopped earlier, and 2) the realized utility of the channel in the j_{th} search is higher than the reservation utility of all the unsearched channels. Thus, the corresponding likelihood function is $L_{ijg}^{Stop}=Pr\left(R_{ijg}\geq u_{ikg},\forall k\in\{1,\ldots,j-1\}\right)\cdot Pr\left(u_{ijg}\geq R_{ilg},\forall l\in\{j+1,\ldots,J\}\right)$. The purchase rule implies that the utility of choosing the last-searched channel is higher than that of the outside option, so the corresponding likelihood function is $L_{ijg}^{Purchase}=Pr\left(u_{ijg}\geq u_{i0g}\right)I(Buy_{ijg}=1)+$

$Pr(u_{ijg} < u_{i0g})\, I(Buy_{ijg} = 0)$. The likelihood of churn (leave permanently) is $L_{ijg}^{Churn} = Pr(\phi_{ijg1} \geq \phi_{ijg0})\, I(Churn_{ijg} = 1) + Pr(\phi_{ijg1} < \phi_{ijg0})\, I(Churn_{ijg} = 0)$. Therefore, the full likelihood is $L = \prod_i \prod_g \prod_j L_{ijg}^{Selection} L_{ijg}^{Stop} L_{ijg}^{Purchase} \left(1 - L_{ijg}^{Churn}\right)$. We use the simulated maximum likelihood approach (SMM) for estimation.

G.3 Identification

Consumers may be enticed to search more under two conditions: an increase in the variance of the utility from search (σ_j) and a reduction in the search cost ($\boldsymbol{\mu}_\gamma$). Thus, σ_j and $\boldsymbol{\mu}_\gamma$ are not identified separately, so we normalize all the $\{\sigma_j\}_{j=1}^{J}$ to 1. The mean utility parameters $\boldsymbol{\mu}_\beta$ are identified by the association between the observed state variables and the frequencies of click and purchase. The heterogeneity utility parameters $\boldsymbol{\Sigma}_\beta$ are identified by the distribution of deviations of the observed clicks and purchases from the predicted values based on the mean utility parameters. The mean cost parameters $\boldsymbol{\mu}_\gamma$ are identified by the inter-consumer observations of continuing or stopping search given their current consideration sets (e.g., corresponding livestream channel positions in the list). The heterogeneity cost parameters $\boldsymbol{\Sigma}_\gamma$ are identified by the deviations from the mean predicted probabilities of stopping. The churn parameters are identified by the association between the state variables and the frequency of churn.

G.4 Result

Table A2 reports the estimated structural parameters of the search model. We include only a subset of the parameters because we use over 300 state variables in the model. The full list of estimates is available from the author upon request.

G.5 Simplification

In the structural sequential search model, the stopping rule and search rule jointly create a mapping between the current state-action pair and the next state because the rules describe whether the consumer is going to churn and, if not, which channel the consumer is going to visit. If we assume that $\mathbf{S}_{it} = \{\mathbf{x}_{it} \backslash \mathbf{A}_{it}\} \cup \mathbf{z}_{it}$, then the stopping rule and search rule can be represented by $\mathbf{S}_{it+1} = g(\mathbf{S}_{it}, \mathbf{A}_{it})$. The purchase rule describes consumer payment, so it can be represented by a mapping between the current state-action pair and the platform's reward; that is, $R_{it} = f(\mathbf{S}_{it}, \mathbf{A}_{it})$. The full structural model characterizes both mappings using the sequential search structure. A simplified model can replace the structural model with flexible functional forms such as GBDT.

Table A2: Model Estimates

	Model With Search				Model Without Search			
	Mean Parameters		Heterogeneity		Mean Parameters		Heterogeneity	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Average price	-0.265	0.027	0.051	0.006	-0.195	0.076	0.043	0.005
Number of viewers (log)	0.592	0.054	0.280	0.072	0.444	0.032	0.234	0.064
Number of likes (log)	0.328	0.033	0.264	0.089	0.360	0.034	0.340	0.077
Coupon Threshold LL Discount LL	0.716	0.210	0.426	0.130	0.568	0.255	0.459	0.101
Coupon Threshold LL Discount LL	1.337	0.406	0.677	0.171	1.198	0.434	0.538	0.212
Coupon Threshold LL Discount M	1.994	0.535	1.107	0.278	1.719	0.573	1.349	0.319
Coupon Threshold LL Discount H	2.594	0.813	1.449	0.428	2.325	0.829	1.785	0.439
Coupon Threshold LL Discount HH	3.285	1.067	1.803	0.503	2.918	1.053	1.943	0.492
Coupon Threshold L Discount LL	0.559	0.170	0.334	0.090	0.488	0.139	0.423	0.074
Coupon Threshold L Discount L	1.306	0.349	0.706	0.189	1.152	0.359	0.681	0.191
Coupon Threshold L Discount M	1.833	0.550	0.965	0.302	1.623	0.522	0.819	0.235
Coupon Threshold L Discount H	2.480	0.740	1.333	0.415	2.205	0.746	1.109	0.356
Coupon Threshold L Discount HH	3.162	1.048	1.848	0.583	2.815	1.036	2.187	0.596
Coupon Threshold M Discount LL	0.525	0.131	0.271	0.074	0.423	0.155	0.320	0.087
Coupon Threshold M Discount L	1.108	0.312	0.633	0.209	0.978	0.306	0.819	0.253
Coupon Threshold M Discount M	1.740	0.475	0.888	0.267	1.559	0.498	0.814	0.260
Coupon Threshold M Discount H	2.494	0.626	1.390	0.363	2.178	0.582	1.344	0.390
Coupon Threshold M Discount HH	3.124	0.824	1.872	0.564	2.775	0.842	1.776	0.476
Coupon Threshold H Discount LL	0.389	0.117	0.200	0.056	0.288	0.155	0.248	0.057
Coupon Threshold H Discount L	1.122	0.313	0.633	0.177	0.966	0.286	0.778	0.210
Coupon Threshold H Discount M	1.756	0.458	1.020	0.264	1.543	0.416	0.747	0.199
Coupon Threshold H Discount H	2.406	0.769	1.206	0.347	2.132	0.768	0.896	0.421
Coupon Threshold H Discount HH	2.977	0.883	1.748	0.570	2.661	0.930	1.948	0.484
Coupon Threshold HH Discount LL	0.257	0.076	0.148	0.043	0.183	0.070	0.192	0.051
Coupon Threshold HH Discount L	0.987	0.294	0.545	0.163	0.824	0.339	0.404	0.198
Coupon Threshold HH Discount M	1.537	0.442	0.899	0.232	1.323	0.453	0.640	0.285
Coupon Threshold HH Discount H	2.191	0.657	1.189	0.386	1.884	0.691	1.046	0.407
Coupon Threshold HH Discount HH	2.987	0.771	1.599	0.402	2.656	0.799	1.879	0.391
Constant	-1.298	0.008	0.405	0.036				
Position	0.152	0.007	0.059	0.018				

Note: The coefficients of the Average price, Number of reviewers (log), Number of likes (log), and Coupon Threshold XX Discount XX are related to the search utility variables $\mathbf{x}_{ijg} = (\mathbf{S}_{it}, \mathbf{A}_{it})'$ where $\mathbf{S}_{it}$ is the vector of the state variables and $\mathbf{A}_{it}$ is the vector of the action dummy variables. The coefficient of the Position variable corresponds to the search cost variable $\mathbf{z}_{\mathbf{ijg}}$ in equation (10).

G.6 Optimization

After estimating the structural parameters, we create an optimization routine to find the optimal policy. The objective of the optimization is to maximize the total discounted revenue over a T period horizon $\max_{A_{it} \in \mathbb{A}} E\left\{\sum_{i=1}^{I} \sum_{t=0}^{T_i} \delta^t R_{it}\left(\mathbf{S}_{it}, A_{it}\right)\right\}$. The finite horizon allows for the dynamic programming problem to be solved via backward induction.

H State Variables and Summary Statistics

Table A3: Static State Variables

Group	State Variables	#
Consumer	Demographics such as age, gender, income, education, and occupation; behavioral variables such as TQZ,[45] purchasing power, product category preference, and host preference.	175
Host	Demographic features of the host as well as the popularity of the host as a seller. Specifically, we include the product categories sold by the host (e.g., jewelry, women's apparel), monthly revenue, a rating of the quality of the host's content (evaluated by experts), and the number of subscribers to the host's channel, etc. The platform pre-defines some host-level summary statistics such as attractiveness, which the platform evaluates from unstructured data, including the host's profile images and voice.	66
Product	Average values for all the products promoted in a livestream,[46] including the product category,[47] price, market share, repurchase rate, review rating, and shipping cost, etc.	36
Livestream (video)	Time of day of the live video (morning, afternoon, or evening); the average number of consumers who watch the livestream, add a product to the cart, write a comment, share the video, like the video, and reward the host; and the total payment amount generated during the livestream, etc. We considered using unstructured data such as textual features (e.g., the video script) and audio features (e.g., the music background), but the data were not available, so we leave it to future research.	27

Table A4: Dynamic State Variables

Group	State Variables	#
Consumer	Related to consumers' coupon reception behaviors: the number of days since the consumer last received a coupon (recency_coupon), the number of coupons received since the beginning of the sample period (frequency_coupon), and the average, minimum, and maximum of the discount ratio and threshold ratio of the coupons received since the beginning of the sample period (monetary_coupon). We also track these variables separately by product category (5 categories: men's/women's/children's apparel, cosmetics, and jewelry).	40
Host	The total number of sellers visited by the consumer since the beginning of the sample period (frequency_seller). Every time a consumer visits a host, she receives a coupon, so the host dynamic state is equivalent to the consumer dynamic state, frequency_coupon. To avoid redundancy, we do not add additional host dynamic state variables.	0
Product	The number of periods (coupon reception incidences) since a consumer last purchased a product (recency_product), the number of products purchased (frequency_product), the average, maximum, and minimum price of the products purchased (monetary_product), and the cumulative spending since the beginning of the sample period (monetary_product). We track these variables separately for each category.	30
Livestream	The number of periods since a consumer last visited a livestream channel (recency_webpage) and the total number of channels visited (frequency_webpage). Every time a consumer visits a channel, she receives a coupon, so the livestream dynamic states are equivalent to the consumer dynamic states, recency_coupon, and frequency_coupon. To avoid redundancy, we do not add additional livestream dynamic state variables.	0
Terminal	The terminal state occurs when the consumer stops returning to the platform. This state captures consumer churn.	1

There are many state variables, so we provide summary statistics for a subset in Table A5, and the complete table is available upon request.

[45] TQZ is an activity score, with higher values indicating higher activity. A consumer can increase her TQZ by searching and buying more products, staying for a longer duration, and writing more reviews.

[46] One livestream session can sell many products, but most of the products usually are in the same category. We use the average product characteristics within a livestream session as state variables.

[47] There are 5 categories: men's apparel, women's apparel, children's apparel, cosmetics, and jewelry.

Table A5: Summary Statistics of State Variables

Category	Static / Dynamic	State Variable	N	Mean	S.D.	Min	25%	50%	75%	Max
Consumer	Static	TQZ	25,886,094	950.58	1406.85	297	664	841	1104	4483
	Dynamic	Frequency coupon	25,886,094	24.35	62.93	0	8	14	26	833
Host	Static	Attractiveness	25,886,094	703.20	900.09	3	532	739	864	999
Product	Static	Price	25,886,094	6067.55	19375.87	12	113	193	320	11298752
Livestream	Static	# Times add to cart	25,886,094	2158.89	6578.47	7	260	810	2397	75451
	Dynamic	Frequency product	25,886,094	1.53	6.04	0	0	0	2	190

I Benchmark Prediction Performance

We evaluate the accuracies of the linear regression, GBDT, and DNN algorithms by splitting the data into training and test sets. Table A6 reports the result. GBDT is the most accurate of the three.

Table A6: Comparison of the Predictive Performance of Three Benchmark Models

		Training Set	Test Set
Number of Consumers		816,718	204,180
Number of Observations		20,688,420	5,197,674
Benchmark Model Mean	Linear Regression	0.93	0.92
	GBDT	0.84	0.89
Squared Error	DNN	0.86	0.91

J Hyperparameters

Table A7: Hyperparameter Choice

Hyperparameter	Definition	Alternatives	Optimal choice
G_ω	Action probability model, Number of trees	5, 10, 50, 100, 500	10
	Action probability model, Maximum depth	1, 2, 5, 10, 20	2
τ	Action probability threshold	0.001, 0.01, 0.05, 0.1, 0.5	0.05
Γ	Target network update rate (number of training iterations)	500, 1k, 8k, 20k	8k
M	Q-network number of nodes in the hidden layers	(16, 16) (32, 32) (64, 64) (100, 50)	(32, 32)
α	Adam optimizer learning rate	0.00001, 0.001, 0.01, 0.1	0.001
ε	Adam optimizer epsilon	0.000001, 0.2	0.2
E	Number of epochs	20, 50	20
N	Minibatch size	256, 512	256

The BCQ algorithm (§4.3) relies on a list of hyperparameters. Table A7 lists the hyperparameters considered, their definitions, value alternatives, and the final optimal values chosen. For instance, for the action probability model G_ω (GBDT), we use 10 trees, each with a maximum depth of 2. The threshold value τ is 0.05. We set the Q-value function approximator as a three-layer, fully-connected neural network. The three layers, in order, have 32, 32, and 1 hidden nodes, and the activation functions are ReLu, ReLu, and linear. The model is trained using the stochastic gradient descent algorithm (step 6 in Algorithm 2) with the Adam optimizer and an epsilon value of 0.2. The mini-batch size is 256, the learning rate is 0.001, the discount factor δ is 0.99, the total number of steps in a trajectory is 25, and the number of epochs is 20. We choose these values by doing a grid search and identifying the hyperparameter combinations with the best

convergence properties.

Following the guideline in Henderson et al. (2018), we use the average return as the evaluation metric to compare different hyperparameters. The two figures below show the average return at each time step for alternative values of two hyperparameters. As shown, our chosen hyperparameter values achieve the highest average return the fastest.

Figure A5: Average Return for Alternative Γ Values

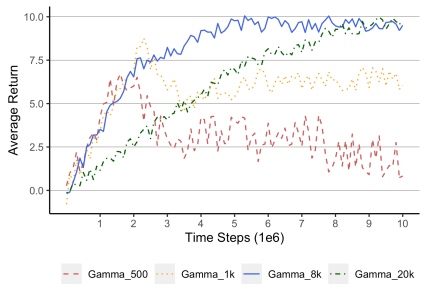
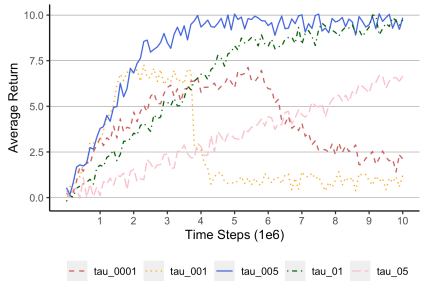


Figure A6: Average Return for Alternative τ Values



We use the XGBoost package (Chen and Guestrin 2016) to estimate GBDT. The number of trees is 200, the learning rate is 0.1, the maximum depth of the individual regression tree is 3, and the fraction of features to consider when looking for the best split is 0.1.

For the neural network model, we use a three-layer, fully-connected neural network with network sizes of 24, 256, and 32 in each layer and ReLu as the activation function. We use the dropout method (dropout rate = 0.2) for regularization.

K Policy Evaluation by Training versus Test Sets

The CLV and gain metrics are slightly lower in the test set than in the training set, but the relative advantage of the proposed BDRL algorithm over the benchmarks is unchanged.

Table A8: Model Comparison Based on the Doubly Robust Estimator

Training Set		1. Static Homogeneous	2. Static Heterogeneous			3. Model-Based Dynamic Heterogeneous	4. Model-Free Dynamic Heterogeneous
		A: Regression	B: GBDT	C: DNN	D: ORF	E: Structural	F: Proposed BDRL
CLV	Mean	6.63	7.64	7.02	7.56	8.41	9.63
(Return)	Std	2.04	2.37	2.30	2.36	2.83	3.27
Gain	Mean	13%	31%	20%	29%	44%	65%
T-test		234.89	499.69	331.93	478.99	646.06	866.16
P_value		< .001	< .001	< .001	< .001	< .001	< .001
Test Set		1. Static Homogeneous	2. Static Heterogeneous			3. Model-Based Dynamic Heterogeneous	4. Model-Free Dynamic Heterogeneous
		A: Regression	B: GBDT	C: DNN	D: ORF	E: Structural	F: Proposed BDRL
CLV	Mean	6.33	7.29	6.88	7.32	8.30	9.33
(Return)	Std	1.88	2.15	2.22	2.27	2.62	3.21
Gain	Mean	6%	23%	16%	23%	39%	57%
T-test		59.36	197.35	134.37	195.05	310.09	392.80
P_value		< .001	< .001	< .001	< .001	< .001	< .001

Note: All the gains are compared to the mean observed CLV: ¥5.85 in the training set and ¥5.95 in the test set. The training set has 816,718 consumers, and the test set has 204,180 consumers.

L Alternative Structural Model Without Search

An alternative structural model can consider only the purchase and churn decisions without the search process (which channel to visit and the sequence of channel visits). For the purchase decision, we consider a random utility-based discrete choice model with random coefficients. The outside option is not purchasing on the livestream platform, and the mean utility is normalized to 0.

$$u_{it1} = \mathbf{x}_{it}^{'}\boldsymbol{\beta}_i + v_{it1}, v_{it1} \sim Gumbel\,(0,1)$$

$$u_{it0} = v_{it0}, v_{it0} \sim Gumbel\,(0,1)$$

$$\boldsymbol{\beta}_i \sim N\left(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta\right)$$

The likelihood of purchase is $L_{it}^{Purchase} = Pr\,(u_{it1} \geq u_{it0})\,I(Buy_{it} = 1) + Pr\,(u_{it1} < u_{it0})\,I(Buy_{it} = 0)$. The churn decision is specified as a discrete choice model.

$$\phi_{it1} = \mathbf{x}_{it}^{'}\rho + \varepsilon_{it1}, \varepsilon_{it1} \sim Gumbel\,(0,1)$$

$$\phi_{it0} = \varepsilon_{it0}, \varepsilon_{it0} \sim Gumbel\,(0,1)$$

The likelihood of churn (leave permanently) is $L_{it}^{Churn} = Pr\,(\phi_{it1} \geq \phi_{it0})\,I(S_{it} = Churn) + Pr\,(\phi_{it1} < \phi_{it0})\,I(S_{it} \neq Churn)$. Therefore, the full likelihood is $L = \prod_i \prod_t L_{ijg}^{Purchase}\left(1 - L_{ijg}^{Churn}\right)$. We use the SMM for estimation. Table A2 reports a subset of the estimated coefficients. The results in Table A9 indicate that consumers in the model without search (vs. in the model with search) are less price-sensitive and less responsive to coupons, a result consistent with prior research (Moraga-González et al. 2015).

Table A9: Model Comparison Based on the Doubly Robust Estimator

		Structural with Search	Structural without Search
CLV	Mean	8.39	8.15
(Return)	Std	2.79	2.52
Gain	Mean	43%	39%
T-test		716.62	657.17
P_value		<0.001	<0.001

Note: All the gains are compared to the mean CLV observed in the data, which is ¥5.87.

M Sensitivity Test for Active Users

To test whether the dynamic pricing recommendations are sensitive to the sample selection rule (users who received at least 10 coupons during the sample period), we split our data into highly-active users (who received more than 25 coupons) and relatively-inactive users (who received 10 to 24 coupons). Below, we provide model-free evidence of the intertemporal tradeoffs within each segment. Fig.A7 and Fig.A8 show similar patterns among the highly-active users and relatively-inactive users. Moreover, we find no behavioral research that suggests that the reference price effect applies only to active users but not inactive ones. Based on the additional data evidence and the absence of relevant research findings, we believe that many of our targeting policy recommendations, learned from active users, will generalize to inactive users.

Figure A7: Current and Future Redemption Rates by Coupon Discount Level and User Activity Level

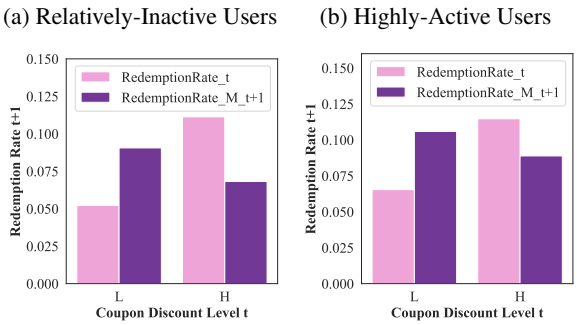
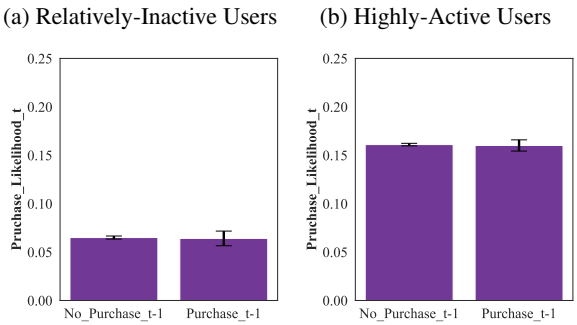


Figure A8: Coupon Redemption Rates by Prior Purchase Experience and User Activity Level



N Policy Evaluation with State Dependence

We consider an alternative specification that can incorporate all three intertemporal tradeoffs discussed in §2.1 and §3.4.2, including state dependence. As shown in Table A10, the monetary value associated with the coupon (monetary_coupon), such as the average/minimum/maximum of the discount ratio/threshold ratio of the coupons received by the consumer, can be considered the reference price, and the product purchase frequency (frequency_product) captures state dependence. A positive state dependence effect indicates inertia, whereas a negative effect indicates variety-seeking.

Table A10: Intertemporal Tradeoffs and Dynamic State Variables

Intertemporal Tradeoffs	Dynamic State Variable
Reference Price	monetary_coupon
Loyalty/Inertia	frequency_product
Variety-Seeking	frequency_product

Table A11 compares the CLV estimates from our model-free dynamic targeting policy (BDRL) and all the benchmark policies. The results are directionally consistent with those in Table 7.

Table A11: Model Comparison Based on the Doubly Robust Estimator

		1. Static Homogeneous	2. Static Heterogeneous			3. Model-Based Dynamic Heterogeneous	4. Model-Free Dynamic Heterogeneous
		A: Regression	B: GBDT	C: DNN	D: ORF	E: Structural[48]	F: Proposed BDRL
CLV	Mean	6.57	7.57	6.99	7.51	8.39	9.57
(Return)	Std	2.01	2.33	2.28	2.34	2.79	3.26
Gain	Mean	+12%	29%	19%	28%	43%	63%
T-test		237.35	536.02	357.17	515.93	716.62	950.56
P_value		<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Note: All the gains are compared to the mean CLV of ¥5.87. The total sample size is 1,020,898 consumers. For separate analyses of the training and test sets, see Appendix K.

O Undiscounted Reward

In our model, the discount factor is not adjusted by the time intervals between consumer visits, which should be the case based on the economic rationale. We choose to use the same discount rate for two reasons. First, the variance in the intervals between coupon reception incidences is relatively small, with an average

[48]The result is for the full structural model. The result from the simplified version is qualitatively similar and available upon request.

inter-incidence time of 3.6 days and a standard deviation of 1.3 days. Second, prior literature that applied reinforcement learning to consumer clickstream data made the same assumption (Urban et al. 2013, Shani et al. 2005, Zheng et al. 2018, and Zou et al. 2019). However, we acknowledge that this assumption is not ideal, so we tested an alternative of undiscounted rewards to alleviate this concern.

Table A12 shows the results with undiscounted rewards. Several interesting results emerge. First, the undiscounted CLVs are larger than the discounted counterparts in Table A11, an expected result. Second, with respect to cross-sectional price discrimination, Fig.A9 shows that the discount factor has minimal impact on the action distributions for less-attractive and attractive hosts. Third, the discount factor has a sizable impact on inter-temporal price discrimination. As Fig.A10 shows, in the undiscounted case relative to the discounted case (Fig.12), BDRL recommends sending fewer high-discount coupon earlier (in the first incidence) and more high-discount coupon later (in the fifth incidence). This may happen because, when future rewards are not discounted, the earlier high-discount coupons have a more pronounced reference price effect, leading the algorithm to recommend fewer high-discount coupons earlier. The discount factor has no effect on the action distribution under the GBDT policy because GBDT is a static policy, not affected by the discount factor. Fourth, the discount ratio time trend in Fig.A11 is steeper in the undiscounted case than in the discounted case (Fig.13). This result is also consistent with the reference price effect; the algorithm reduces the discount ratio earlier and increases the discount ratio later to avoid the negative reference price effect. Overall, the results in the undiscounted case are qualitatively similar to those reported in the main text.

Table A12: Model Comparison with Undiscounted Rewards Based on the Doubly Robust Estimator

		1. Static Homogeneous	2. Static Heterogeneous			3. Model-Based Dynamic Heterogeneous	4. Model-Free Dynamic Heterogeneous
		A: Regression	B: GBDT	C: DNN	D: ORF	E: Structural[49]	F: Proposed BDRL
CLV	Mean	8.01	9.22	8.53	9.13	10.21	11.66
(Return)	Std	2.48	2.85	2.78	2.86	3.42	3.99
Gain	Mean	11%	28%	19%	27%	42%	62%
T-test		261.88	582.36	393.52	555.66	761.54	1,001.05
P_value		<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

[49]The result is for the full structural model. The result from the simplified version is qualitatively similar and available upon request.

Figure A9: Targeting Rule Under BDRL (Undiscounted Case): Host Attractiveness

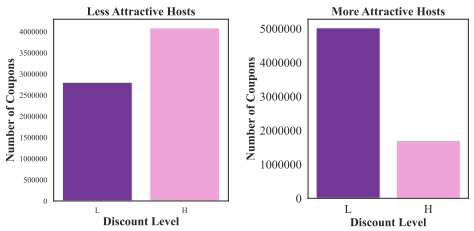
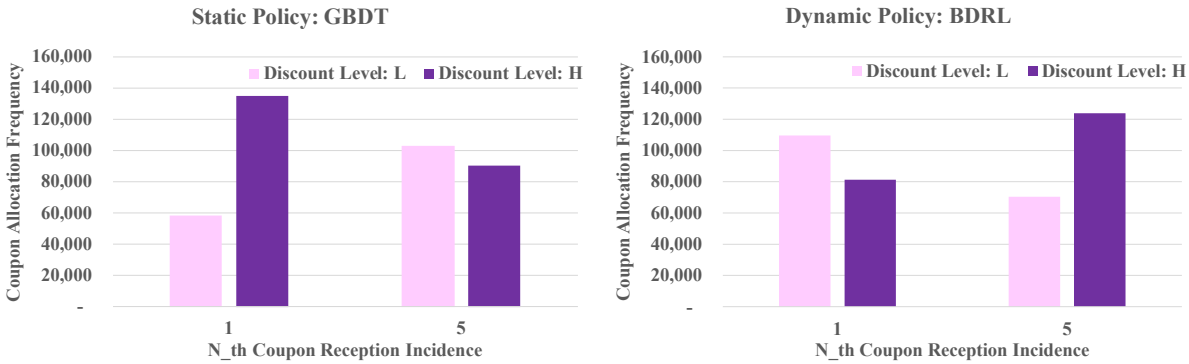
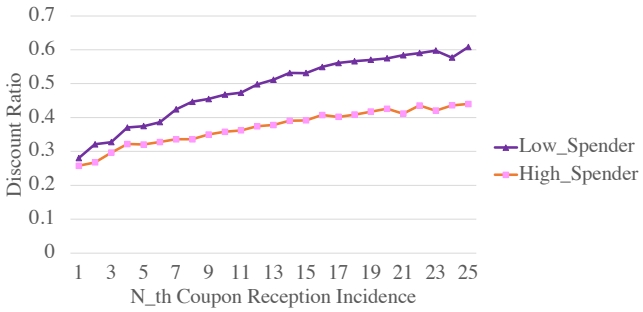


Figure A10: Comparison Between the Static and Dynamic Targeting Policies (Undiscounted Case)



O.0.1 Cross-sectional and Intertemporal Price Discrimination

Figure A11: Targeting Rule (Undiscounted Case): Intertemporal Price Discrimination



P Model Comparison Using Distributional Differences

We use the Kolmogorov-Smirnov test to measure distributional differences. As shown below, both the Doubly Robust Estimator and field experiment results are consistent with the t-test results reported previously.

Table A13: Model Comparison Based on the Doubly Robust Estimator (KS test)

		1. Static Homogeneous	2. Static Heterogeneous			3. Model-Based Dynamic Heterogeneous	4. Model-Free Dynamic Heterogeneous
		A: Regression	B: GBDT	C: DNN	D: ORF	E: Structural[50]	F: Proposed BDRL
CLV	Mean	6.57	7.57	6.99	7.51	8.39	9.57
(Return)	Std	2.01	2.33	2.28	2.34	2.79	3.26
Gain	Mean	+12%	29%	19%	28%	43%	63%
KS-test		0.14	0.31	0.21	0.30	0.42	0.54
P_value		<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Table A14: Field Experiment Results (KS test)

		Random Allocation	Model-Based Dynamic Heterogeneous (Structural)	Model-Free Dynamic Heterogeneous (BDRL)
CLV	Mean	6.98	9.70	11.16
(Return)	Std	2.43	3.15	3.86
Gain	Mean	~	+39%	+60%
KS-test		~	0.38	0.52
P_value		~	<0.001	<0.001

Q State Transition Implementation in Different Algorithms

The state transition process described in §4.2.2 is implemented differently in the three algorithms (Q-learning, BCQ, and the structural models).

In Q-learning, state transitions are observed by the agent. As shown in Algorithm 1, once the agent takes action A, the environment returns R and S' to the agent (line 7). In our livestream setting, if the firm deploys Q-learning online (in real time), then after the firm takes action A, the consumer will make the purchase, search, and churn decisions, and these consumer decisions will result in R and S'. So, the firm can observe the next state when updating the Q function. There is no need to use historical data to estimate the state transition matrix.

In BCQ, state transitions are observed from the batch data. As shown in Algorithm 2, a sample of M transitions (S,A,R,S') is drawn from the batch data in step 4. An action is chosen in step 5, and the Q function updates in step 6 (specifically, the θ parameters in the neural network approximation of the Q

[50]The result is for the full structural model. The result from the simplified version is qualitatively similar and available upon request.

function). Also, the state transition S' is observed from the sampled transitions in step 6. Again, there is no need to use historical data to estimate the state transition matrix.

In the structural model, state transitions are predicted using the structural parameters. Specifically, in the value function iteration step (equation 8), the expectation is taken over the state transition probabilities, and we rely on the estimated state transition to calculate the expectation. As explained in §4.2.2, state transitions can be fixed, stationary, or stochastic. For the static state variables that follow a stationary distribution, in each coupon reception incidence t, we randomly draw a value from the stationary distribution of the state variables. The multivariate (joint) state variable distribution is estimated prior to training the targeting policy. The transitions of the dynamic (stochastic) state variables are governed by the structural model. For instance, the transitions of product-related dynamic variables can be calculated based on Table 5 once Purchase is predicted using the structural model.

References

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD*, pages 785–794, 2016.

James J Heckman. 3. heterogeneity and state dependence. In *Studies in labor markets*, pages 91–140. University of Chicago Press, 2007.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Paul Klemperer. The competitiveness of markets with switching costs. *The RAND Journal of Economics*, pages 138–150, 1987.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994*, pages 157–163. Elsevier, 1994.

José L Moraga-González, Zsolt Sándor, and Matthijs R Wildenbeest. Consumer search and prices in the automobile market. 2015.

Scott A Neslin, Caroline Henderson, and John Quelch. Consumer promotions and the acceleration of product purchases. *Marketing Science*, 4(2):147–165, 1985.

PB Seetharaman and Hai Che. Price competition in markets with consumer variety seeking. *Marketing Science*, 28(3):516–525, 2009.

Guy Shani, David Heckerman, Ronen I Brafman, and Craig Boutilier. An mdp-based recommender system. *Journal of Machine Learning Research*, 6(9), 2005.

Xuanming Su. Intertemporal pricing with strategic customer behavior. *Management Science*, 53(5):726–741, 2007.

Shining Wu, Qian Liu, and Rachel Q Zhang. The reference effects on a retailer's dynamic pricing and inventory strategies with strategic consumers. *Operations Research*, 63(6):1320–1335, 2015.

Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. Drn: A deep reinforcement learning framework for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*, pages 167–176, 2018.

Lixin Zou, Long Xia, Zhuoye Ding, Jiaxing Song, Weidong Liu, and Dawei Yin. Reinforcement learning to optimize long-term user engagement in recommender systems. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2810–2818, 2019.