Sequential Search with Refinement: Model and Application with Click-stream Data

Yuxin Chen Song Yao [1]

September 26, 2014

[1]Yuxin Chen is the Distinguished Global Professor of Business, NYU Shanghai with affiliation with the Stern School of Business, New York University (email: yc18@nyu.edu). Song Yao is an Assistant Professor of Marketing at the Kellogg School of Management, Northwestern University (email: s-yao@kellogg.northwestern.edu). This project was previously circulated under the working title "Search with Refinement." The authors would like to thank seminar participants at Cornell University, Hong Kong University of Science and Technology, Ohio State University, University of Chicago, Washington University at St. Louis, Marketing Science Conference 2012, Third Annual Searle Conference on Internet Search and Innovation, INFORMS International Conference 2012, as well as Paulo Albuquerque, Bart J. Bronnenberg, Xinlei (Jack) Chen, Pradeep Chintagunta, Anindya Ghose, Günter Hitsch, Jun Kim, Dmitri Kuksov, Carl Mela, Chun-Hui Miao, Sridhar Moorthy, Harikesh Nair, and Ting Zhu for their feedback. The authors thank an anonymous travel website and Wharton Customer Analytics Initiative (WCAI) for providing the data.

Abstract: Sequential Search with Refinement: Model and Application with Click-stream Data

We propose a structural model of consumer sequential search under uncertainty about attribute levels of products. Our identification of the search model relies on exclusion restriction variables that separate consumer utility and search cost. Because such exclusion restrictions are often available in online click-stream data, the identification and corresponding estimation strategy is generalizable for many online shopping websites where such data can be easily collected. Furthermore, one important feature of online search technology is that it gives consumers the ability to refine search results using tools such as sorting and filtering based on product attributes. The proposed model can integrate consumers' decisions of search and refinement. The model is instantiated using consumer click-stream data of online hotel bookings provided by a travel website. The results show that refinement tools have significant effects on consumer behavior and market structure. We find that the refinement tools encourage 34% more searches and enhance the utility of purchased products by 18%. However, most websites by default rank search results according to their qualities or relevance to consumers (e.g., Google). When consumers are unaware of such default ranking rules, they may engage in disproportionately more searches using refinement tools. Consequently, overall consumer surplus may deteriorate when search cost outweighs the enhanced utility. In contrast, if the website simply informs consumers that the default ranking already reflects product quality or relevance, consumers search less and their surplus improves. We also find that refinement tools lead to a less concentrated market structure.

Keywords: consumer search, click-stream data analysis, electronic commerce, consumer behavior

1 Introduction

According to a recent report by McKinsey, the annual global value of search technology is $780 billion with $540 billion direct contribution to global GDP.[1] The advance of online search technology has made profound impacts on consumer behavior. In particular, search technology has helped consumers to easily form a consideration set among many products with unknown attribute levels. Because consumers' final purchase decisions depend on their consideration sets, understanding how consumers engage in such searches has become crucial for firms.

However, measuring consumer search and purchase activities using field data remains difficult. For example, because researchers do not observe consumer preference or search cost in the data, a consumer's decision to stop the search can be attributed to either low preference of the next search or high search cost (Sorensen, 2000; Koulayev, 2013a). We propose an identification and estimation strategy of a sequential search model that relies on exclusion restrictions to separate consumer preference and search cost. Such exclusion restrictions are presented in many click-stream data of shopping websites. Upon the separation of consumer preference and search cost, the identification of the model works similar to classical discrete choice models. The model can be applied to the situation where products have multiple attributes and consumers are uncertain about the attribute levels of unsearched products. More importantly, our proposed model is generalizable for many shopping websites that have access to click-stream data.

This paper advances the growing empirical literature on identifying search models. Hong and Shum (2006) and Hortacsu and Syverson (2004) develop structural approaches to estimate the distribution of consumer search costs using aggregate data. Their approaches utilize parameter restrictions implied by equilibrium conditions such as equilibrium price distribution derived from the supply side. Moraga-Gonzalez et al. (2012) also use aggregate data and estimate a simultaneous search model (search with fixed consideration set size)

[1] "The Impact of Internet Technologies: Search," July 2011, McKinsey.

in the context of the automobile market. The search cost distribution is recovered using exogenous variations observed in the market (distances to dealerships). De los Santos et al. (2012) use individual-level comScore data on web-browsing and purchase to explore which classical search model is more consistent with observed data patterns. Studies by Koulayev (2013a) and Honka and Chintagunta (2013) come closest to our proposed model. In Koulayev (2013a), the author observes click-stream data on search but not purchase activities. Because the data contain the sequence of search activities, the identification relies on (1) the binary decision to continue or stop the search, and (2) the variation of attributes among previously searched products. We propose to use exclusion restrictions to separate the preference and the search cost. Moreover, because purchase data are also commonly available to shopping websites, when combined with search data, they help to better identify consumer preference. This is because that, conditioned on the consideration set, a consumer's purchase decision is only subject to her preference but not her search cost. Honka and Chintagunta (2013) identify consumers' search methods (sequential vs. simultaneous search) in the context of price searches of alternative automobile insurance plans. The data contain individual-level information on consumers' consideration sets and final purchases. However, the researchers do not observe search sequences. The authors propose using price patterns in the observed consideration sets to help identify search methods. In comparison, we focus on sequential search and utilize observed search sequences, further accessing search activities for attributes in addition to price and better controlling for unobserved heterogeneity.

Furthermore, modern search technology allows users who search products or services with multiple attributes to refine search results. For example, an individual searching for a hotel on a travel website may sort results by price in an ascending order and filter out hotels with star ratings below three. As another example, an academic researcher may conduct a keyword search for journal articles in an online library, filtering out all non-peer-reviewed titles and sorting results by publication dates. Given the ubiquitous use of such refinement tools in online search, it is surprising that few empirical studies focusing on

the refinement tools' value to consumers and their impact on consumer search behavior and market structure. Accordingly, another objective of this paper is to fill this gap. It is possible for our model to incorporate consumer sorting and filtering on multiple product/service attributes. Specifically, in our model, consumers may apply refinement tools to alter the distribution of attributes. Our empirical findings and subsequent counterfactual analyses suggest that, with the aid of refinement tools, consumers' searches increase by about 34%. Furthermore, on average a consumer may achieve 18% higher utility for the product she chooses. It is crucial that websites educate consumers about their practice of ranking search results according to their qualities or relevance to consumers (e.g., Google). We find that consumers engage in disproportionately excessive searches using refinement tools when they are uninformed about such default ranking rules. The cost accrued during the search process outweighs the utility improvement of the purchased product. Consequently, uneducated consumers' overall surplus actually drops when they use refinement tools. The welfare loss due to excessive searches can amount to -3.4%. In contrast, when consumers understand that the default ranking is according to qualities, they search less and their surplus exceeds the level without refinement tools. We further consider a possible alternative ranking rule for search results. We show that by incorporating our model results into the ranking rule, consumer welfare can be further improved by 1.6%. This is consistent with the findings of Ghose et al. (2012) and Ghose et al. (2014), which show that consumer surplus improves when the ranking incorporates consumer utility information. We also find that the market becomes less concentrated owing to the existence of refinement tools because heterogeneous consumers are able to locate differentiated hotels that match their preferences better. They search with greater depth and find more hotels. Such better matches would be too costly to achieve without the refinement tools.

In addition, this paper extends the empirical literature in marketing and economics on consideration set formation. Mehta et al. (2003), Kim et al. (2010), Honka (2012), Seiler (2013), and Chan and Park (2014) propose structural models for the formation of considera-

tion sets as the result of consumer search, and model consumer purchase conditional on the consideration sets. While Kim et al. (2010) adopt sequential search assumption, Mehta et al. (2003), Seiler (2013), and Honka (2012) consider simultaneous search assumption. Chan and Park (2014) consider the context of sponsored search advertising and emphasize advertisers' perspective. Accordingly, they adopt a simplified model where consumers search in the order of slot positions on the webpage. The actual search process and search behavior are not observed in the studies mentioned, with the exceptions of Honka (2012) and Chan and Park (2014). A distinguishing feature of Honka (2012) and Chan and Park (2014) is their observation of consumers' consideration sets. However, neither datasets contain information of search sequences. Chan and Park (2014) also does not observe purchase activities. In contrast, we focus on the online shopping context where click-stream data are routinely collected by the firms. Taking advantage of the availability of consideration set, search sequence, and purchase information in these click-stream data, we consider the sequential search and purchase decisions (including the usage of search refinement tools). This enables us to build a structural model in which consumer decisions on search, refinement, and purchase are derived from a unified framework of utility maximization.

Our paper is also related to Yao and Mela (2011), which explicitly models consumer decisions to use sorting and/or filtering functions in online search. Their model is constructed from the perspective of online advertisers. To be consistent with the information structure of the advertisers, the model aggregates individual consumer choices up to the market level. In contrast, our model addresses the search at the individual consumer level, enabling us to address subtle issues, such as how refinement affects the number of searches.

The rest of the paper is organized as follows. In Section 2, we detail the structural model of consumer optimal sequential search using online click-stream data. We then describe the estimation approach. In Section 4, we discuss the identification and present some Monte Carlo simulation results. We use a click-stream dataset to demonstrate the application of our model in Section 5, where we also present several counterfactual simulations to explore

the managerial implications. We conclude with a discussion of main findings and suggestions for future research.

2 A Model of Sequential Search with Online Click-stream Data

2.1 Online Shopping Click-stream Data

Click-stream data are commonly available to online shopping websites. Such data normally contain information on individual consumers' click-throughs and purchases. At the same time, the websites have access to product information such as product attribute levels, promotions, and website design information such as slot positions of products on their webpages. We specify the model in a general framework that captures the main features of these data.

When a consumer arrives at an online shopping website with the intention to purchase certain product (e.g., hotel in our application later), the website presents a list of products for the consumer to consider. The products are positioned at different slots on the webpage. For a given slot, the consumer is uncertain about product attribute levels and hence her utility level before clicking through the link. Accordingly, we define a search as the consumer clicking through the link at a slot position. The click-through resolves the uncertainty about her utility level. This definition of search is consistent with the classical economic literature (e.g., Nelson, 1970).[2]

We assume consumers engaging in sequential search. Before the search starts, the consumer knows the utility level of the outside option of not buying. Let j be the index of the sequence in which consumer i searches. At a given point during the search, suppose that the consumer has already searched $j-1$ slots, the consumer needs to decide (1) whether she should continue with the j-th search, (2) if yes, which slot position to search, (3) if no, which searched product to purchase (including the outside option). These decisions depend on the tradeoff between one's utility and search cost.

[2]We use "search" and "click-through" interchangeably henceforth.

2.2 Utility

Consumer i's utility of buying the product searched during the j-th search is characterized as

$$u_{ij} = \mu_i(x'_{ij}) + \nu_{ij}, \tag{1}$$

where x_{ij} is a vector of product attributes. x_{ij} may vary across consumers. For example, product attribute levels, such as price, may depend on when the consumer makes the inquiry. ν_{ij} is a consumer-product specific idiosyncratic preference shock. x_{ij} is drawn from some joint distribution of product attributes, $P_j(x_{ij})$. ν_{ij} is i.i.d. across consumers and products.

2.3 Search Cost

The consumer incurs a search cost for each additional search. The search cost can be interpreted as time and efforts spent on the search. Denote the search cost as:

$$c_{ij} = c_i(z'_{ij}) \tag{2}$$

where z_{ij} is a vector of consumer and search related characteristics. The search cost may depend on some consumer characteristics. For example, consumers who are more time-constrained may be subject to a higher level of search cost (McDevitt, 2013). The search cost also depends on some characteristics of the particular search. For example, the search involves different slot positions on the webpage. Because slot positions may affect the accessibility, the search cost may depend on the slot position of that search (Ansari and Mela (2003); Yao and Mela (2011)).

With these specifications, we now formalize what the consumer does (and does not) know *before* the j-th search:

- The consumer knows the search cost for the j-th search, c_{ij}.

- The consumer knows the distribution of product attributes $P_j(x_{ij})$.
- The consumer knows the distribution of ν_{ij}.
- The consumer has a rational expectation about the expected utility of the search.

2.4 Expected Marginal Gain of an Additional Search

Suppose the consumer has already made $j-1$ searches. Denote u_i^* as the maximum utility among those searched options. For the next search, the j-th search, denote the CDF of u_{ij} as $F(u_{ij})$. The distribution $F(u_{ij})$ depends on the distribution of ν_{ij} and the distribution of x_{ij}, $P_j(x_{ij})$. The expected marginal net gain from making the j-th search, then stopping the search and choosing the option with the highest utility is given by Weitzman (1979) as

$$Q_{ij} = \int_{u_i^*}^{\infty} (u_{ij} - u_i^*) dF(u_{ij}) - c_{ij} \tag{3}$$

where the integral is the expected improvement in utility if u_{ij} is greater than u_i^*. "$-c_{ij}$" signifies that the consumer needs to pay the search cost for the j-th search.

2.5 Optimal Sequential Search Strategy

Before characterizing the optimal search strategy, we first define the consumer's reservation utility R_{ij}, which is the utility level that makes the consumer indifferent between (1) choosing an already-searched option with the utility level of R_{ij}^k, and (2) making the j-th search. That is, R_{ij} solves the implicit function

$$Q_{ij} = \int_{R_{ij}}^{\infty} (u_{ij} - R_{ij}) dF(u_{ij}) - c_{ij} = 0 \tag{4}$$

As shown in Weitzman (1979), given c_{ij} and $F(u_{ij})$, a unique reservation utility R_{ij} solves equation 4. And the optimal search strategy contains two steps: a stopping rule to determine when to stop searching and a selection rule for how to search.

Step 1: Stopping Rule (when to stop searching): Calculate the reservation utility for each alternative search option. If no reservation utility exceeds the then-current maximum utility u_i^*, stop the search and choose the searched option with the highest utility u_i^*. Otherwise, proceed to the next step.

Step 2: Selection Rule (how to search): Search the alternative with the highest reservation utility, update u_i^*, and go back to Step 1.

This optimal strategy can be interpreted as follows: The consumer will continue searching if the expected marginal gain is positive. In particular, she will choose to search the option with the highest reservation utility. If the consumer decides to stop the search, then she will choose the searched option with the highest utility among those already searched.

3 Estimation

In this section, we detail the estimation strategy for the model described above.

3.1 Heterogeneous Utility

We assume that preference heterogeneity enters the utility function in a linear fashion such that:

$$\begin{aligned} u_{ij} &= \mu_i(x'_{ij}) + v_{ij} \\ &= x'_{ij}\alpha_i + v_{ij} \\ &= x'_{ij} \cdot (\alpha + \xi_i \odot \sigma_u) + \nu_{ij} \end{aligned} \tag{5}$$

where x_{ij} is the vector of product attributes. α is a column vector of the averages of consumers' sensitivities pertaining to product attributes. ξ_i and σ_u are column vectors with the same dimension as α. The notation "$\odot$" stands for element-wise multiplication of two vectors of the same dimension. $\xi_i \odot \sigma_u$ represents individual heterogeneity in preference, measuring individual i's deviations from the average sensitivities α. Let individual preference heterogeneity follows some known distribution. In particular, we assume that ξ_i follows standard

normal distribution, $\xi_i \sim N(0, I)$. σ_u then captures the magnitude of the heterogeneity. The random error term v_{ij} follows a standard normal distribution and is i.i.d. across individuals and products.

3.2 Heterogeneity in Search Cost

To capture the heterogeneity in search cost, we specify search cost as

$$\begin{aligned} c_{ij} &= c_i(z'_{ij}) \\ &= z'_{ij}\gamma_i \\ &= z'_{ij} \cdot (\gamma + \zeta_i \odot \sigma_c) \end{aligned} \tag{6}$$

where γ are mean levels of cost coefficients. $\zeta_i \odot \sigma_c$ are the deviations from the mean levels and capture the heterogeneity in search cost. ζ_i follows standard normal distribution such that $\zeta_i \sim N(0, I)$ and σ_c measures the magnitude of the heterogeneity.

3.3 Likelihood

We can write down the likelihood function based on the optimal search strategy described in Section 2.5. However, one complication comes from the dependency between purchase and search decisions: the purchase is conditioned on the consideration set, which is endogenously determined by the search process. Accordingly, we consider a simulated maximum likelihood approach that accounts for the dependency. We describe the two components of the likelihood function next.

3.3.1 Purchase Likelihood

The utility of choosing product j is specified as:

$$u_{ij} = x'_{ij} \cdot (\alpha + \xi_i \odot \sigma_u) + \nu_{ij}$$

In purchase data, we observe the consideration set of each consumer, i.e., the products searched by each consumer. We also observe each consumer's final purchase. Denote S_i as the consideration set of consumer i, containing all products searched by consumer i. Let j^* be the final purchase by i. We have the corresponding purchase likelihood as

$$\begin{aligned} u_{ij^*} &\geq u_{ij}, \forall j \in S_i \\ L_i^{purchase} &= \Pr(u_{ij^*} \geq u_{ij}, \forall j \in S_i) \\ &= \prod_{\forall j \in S_i} \Pr(u_{ij^*} \geq u_{ij}) \\ &= \prod_{\forall j \in S_i} \Pr(x'_{ij^*}\alpha + x'_{ij^*}\xi_i \odot \sigma_u + \nu_{ij^*} \geq x'_{ij}\alpha + x'_{ij}\xi_i \odot \sigma_u + \nu_{ij}) \end{aligned} \tag{7}$$

3.3.2 Search Likelihood

Denote S_{ij} as the set of products searched before the j-th search. According to the optimal search strategy, when the consumer engages in the j-th search, the reservation utility of that search option exceeds the realized utilities of all searched products. The corresponding likelihood function is

$$\begin{aligned} L_i^{search} =& \Pr(R_{ij} \geq u_{ir}, \forall r \in S_{ij}) \\ &= \prod_{\forall r \in S_{ij}} \Pr(R_{ij} \geq u_{ir}) \\ &= \prod_{\forall r \in S_{ij}} \Pr(R_{ij} \geq x'_{ir}\alpha + x'_{ir}\xi_i \odot \sigma_u + \nu_{ir}) \\ &= \prod_{\forall r \in S_{ij}} \Pr(R_{ij} - (x'_{ir}\alpha + x'_{ir}\xi_i \odot \sigma_u) \geq \nu_{ir}) \end{aligned} \tag{8}$$

3.3.3 Joint Likelihood

The joint likelihood for all consumers is therefore

$$L = \prod_i L_i^{purchase} L_i^{search} \tag{9}$$

However, the last lines in Equation 7 and Equation 8 do not have closed form solutions. This is because both heterogeneity and revealed preference shocks of searched products (ν_{ij}) are known to consumers but not observed by the researchers. They need to be integrated out when we construct the likelihood function. More importantly, the distribution of ν_{ij} for those already-searched products is truncated from the perspective of the researchers. To be specific,

1. Denote the final search as $\bar{j}$. For any search before the final one, $\forall j < \bar{j}$, we may conclude that the reservation utilities of those searches after j are greater than u_{ij} because the search continues afterwards. Hence we have

$$\begin{aligned} u_{ij} &= x'_{ij}\alpha_i + v_{ij} < \min(R_{ij'}, \forall j < j' \leq \bar{j}) \\ v_{ij} &< \min(R_{ij'}, \forall j < j' \leq \bar{j}) - x'_{ij}\alpha - x'_{ij}\xi_i \odot \sigma_u \end{aligned}$$

2. Because the search stops after the final search $\bar{j}$, we may conclude that at least the purchased product has a utility level greater than all remaining unsearched reservation utilities. Hence we have

$$\begin{aligned} u_{ij^*} &= x'_{ij^*}\alpha_i + v_{ij^*} > \max\{R_{i\tilde{j}}, \forall \tilde{j} \in \widetilde{S}_i\} \\ v_{ij^*} &> \max\{R_{i\tilde{j}}, \forall \tilde{j} \in \widetilde{S}_i\} - x'_{ij^*}\alpha - x'_{ij^*}\xi_i \odot \sigma_u \end{aligned}$$

As a result, ν_{ij}'s of searched products are no longer normally distributed. From the perspective of the researchers, for searched products that are not purchased, the preference shocks

are right truncated. For the purchased product, the shock is left truncated if it is the final search; if it is not the final search, it is truncated on both sides.

Because no closed form solutions exist for the probabilities, we propose a simulated method to construct the estimable likelihood function. To be specific,

1. Given $\xi_i \sim N(0, I)$ and $\zeta_i \sim N(0, I)$, make one random draw of the pair (ξ_i, ζ_i) for consumer i.

2. Conditional on a given set of parameters, and the pair of (ξ_i, ζ_i), for each product (x_{ij}) in the consideration set, draw 100 ν_{ij}, depending on the truncation conditions aforementioned.

3. Calculate the frequency of the condition $\{u_{ij^*} \geq u_{ij}, \forall j \in S_i\}$ being satisfied across the random draws of ν_{ij}'s.

4. Calculate the frequency of the condition $\{R_{ij} - (x'_{ir}\alpha + x'_{ir}\xi_i \odot \sigma_u) \geq \nu_{ir}, \forall r \in S_{ij}\}$ being satisfied across the random draws of ν_{ij}'s.

5. Repeat Step 1-4 for 100 times, repeatedly making new pairs of draws for (ξ_i, ζ_i). The average of the simulated probabilities from Step 3 and Step 4 across these 100 pairs of (ξ_i, ζ_i) is the simulated probabilities for consumer i.

6. Repeat Step 1-5 for all consumers to obtain the simulated likelihood for Equation 9.

3.3.4 Calculation of the Reservation Utilities

To speed up the estimation, we follow Kim et al. (2010) to impute R_{ij} outside the estimation loop.[3] It can be shown that the following equation holds:

$$c_{ij} = \{(1 - \Phi(R_{ij} - \mu_{ij}))(\mu_{ij} - R_{ij} + \frac{\phi(R_{ij} - \mu_{ij})}{(1 - \Phi(R_{ij} - \mu_{ij}))})\} \tag{10}$$

[3]In comparison to the approach used in Kim et al. (2010), we further consider the uncertainty of attributes x_{ij}.

where $\mu_{ij} = x'_{ij}\alpha_i$, i.e., the expected utility level with the preference shock ν_{ij} integrated out. One subtle but important point is that Equation 10 relies on the shock ν_{ij} being normally distributed. Unsearched ν_{ij} is unknown to both the consumer and the researchers. Consequently, when the consumer evaluates the reservation utility of an unsearched option, ν_{ij} does follow normal distribution. In contrast, searched ν_{ij} is known to the consumer but unknown to the researchers. So the distribution becomes truncated from the researchers' perspective as mentioned above.

For a given pair of $\{c_{ij}, \mu_{ij}\}$, we can calculate the corresponding reservation utility R_{ij} by solving Equation 10. We can simplify the computation by constructing a look-up table of the triple $\{c, \mu, R\}$, with the grid up to a substantial fine level. Since the table holds for all searches, we drop the subscripts. Note that this grid does not depend on the parameter values. We can first create this table outside of the estimation loop. Then during the estimation for each given pair of $\{c_{ij}, \mu_{ij}\}$, we use the table to impute the corresponding value of R_{ij}, potentially with an interpolation step if the table does not contain the exact pair of $\{c_{ij}, \mu_{ij}\}$.[4] We address the uncertainty of x_{ij} and hence μ_{ij} by making draws repeatedly from the attributes distribution and then calculating the corresponding expectation of R_{ij}.

4 Identification and Monte Carlo Simulations

4.1 Identification

4.1.1 Separating Utility and Search Cost

The identification of search model using field data is difficult because the interdependence between the search cost and the preference.[5] In standard discrete choice models, utility parameters can be identified from purchase data alone. Two necessary conditions for the identification are (1) one alternative's utility level or one attribute's coefficient is normalized

[4] We use a third-order polynomial regression in our implementation.

[5] Consumers' beliefs about the distribution of attributes may further confound the identification. In scenarios where consumers are familiar with the products, we may assume the consumer knows the attributes or the distribution of attributes and has rational expectation regarding the marginal gain of the search (see Section 2.3). We call for future research to relax such an assumption so that consumers learn the distribution during the search.

(e.g., outside product has mean zero utility or price coefficient is -1), and (2) the distribution of preference shocks is assumed (e.g., standard normal or logistic distribution). In the focal setting, we may use similar normalization to satisfy the first condition. It is, however, trickier to satisfy the second condition. Those preference shocks (ν_{ij}) within a consumer's consideration set (after the search) have truncated distributions to the researchers; and the truncation depends on the search cost (see Section 3.3.1).

Correspondingly, we consider exclusion restrictions for separating the utility and the search cost, an identification strategy similar to classical selection models. Purchase decision is based on the utility specified in Equation 1:

$$u_{ij} = \mu_i(x'_{ij}) + \nu_{ij}$$

We observe purchase data, i.e., conditional on the consideration set and the truncations of preference shocks, consumers' decisions on which products to buy. Such a setting is similar to the "outcome equation" in selection models. As for the "selection equation," it depends on both the utility function as well as the search cost specified in Equation 2:

$$c_{ij} = c_i(z'_{ij})$$

The decision rule that determines the formation of the consideration set is according to the implicit function of reservation (Equation 4):

$$\int_{R_{ij}}^{\infty} (u_{ij} - R_{ij}) dF(u_{ij}) - c_{ij} = 0$$

When we choose different sets of covariates for x_{ij} and z_{ij}, the covariates enter search cost function but not utility function serve as the exclusion restrictions for identification. In addition, because the implicit function is nonlinear, the nonlinearity further helps the separation between utility and search cost.

Conditioned on the exclusion restrictions such that the utility and the search cost can be separated, the identification of preference and search cost parameters are similar to classical discrete choice models. We next discuss the identification of preference and search cost parameters, respectively.

4.1.2 Identifying Preference Parameters

Mean preference parameters are identified from both purchase data and search data.

- In purchase data, we observe (1) product attributes x_{ij} within each consumer's consideration set, and (2) the final choice of each consumer conditioned on her consideration set. The final purchases given x_{ij} across consumers and products reveal the mean levels of preference parameters similar to classical multinomial discrete choice models.

- In search data, before the j-th search, we observe: (1) attributes of products up to the j-th search (2) product attributes (exact levels or the distribution) of the j-th search. Across consumers, given what have been searched in each's consideration set, the next search's attributes further help the identification of mean preference parameters. For example, if on average people tend to pick high quality and high price options for the next search, we may conclude that people have a low price sensitivity and care more about quality. To some extent, for each search, it is similar to a binary choice model where the consumer has two options: make the j-th search or not. The difference is that the "baseline utility" in a standard binary choice model is normalized to zero. In contrast, it changes over the course of the search in the focal sequential search setting.

The identification of preference heterogeneity relies on both purchase data and search data. Preference heterogeneity σ_u cannot be easily recovered based on purchase data alone. This is because that one common feature for click-stream data is the sparsity of repeated purchase. For most consumers, we only observe one purchase incidence per consumer. In contrast, across both purchase data and search data, we have multiple observations per consumer. For a given consumer and her search cost, we observe the deviation of observed purchase

and searches from those predicted decisions based on mean preference parameters. The distribution of these deviations across individual consumers identifies the heterogeneity distribution parameters σ_u.

4.1.3 Identifying Search Cost

From search data, we observe the consideration sets, sequences of search, and z_{ij}'s across consumers. Across consumers and their searches, conditioned on the preference and search cost can be separated due to the exclusion restrictions, we observe average proportions of consumers continuing or stopping the search given their then-current consideration sets and unsearched options. It is clear from the marginal gain of each search (Equation 3) that the consumer is essentially trading off (1) then-current maximum utility among the searched, and (2) expected utility of the search net the search cost. These across consumers observations of continuing or stopping search given their then-current consideration sets identifies mean search cost among consumers. Furthermore, at each point during a given consumer's search, based on mean parameters, her z_{ij}, and the products already searched before her j-th search, we may predict the mean probability of her stopping the search. The deviation of her search activities from these predicted values give us the information of one's heterogeneity in search cost. The distribution of these deviations across individual consumers identifies the heterogeneity distribution parameters σ_c.

4.2 Monte Carlo Simulations

We use Monte Carlo simulations to demonstrate the feasibility of model identification. In particular, if search cost and utility can be empirically separated, the identification of preference and search cost parameters is relatively standard. So we focus the simulations on separating the search cost and the utility using exclusion restrictions.

We use the search model detailed in Section 2 to simulate four datasets. The first two datasets each has 200 consumers and 100 products. The utility of a consumer for a product depends on the product's price, quality, a baseline utility (constant term), and a random

preference shock. The true coefficients of price, quality, and the constant are -2, 2, and 5, respectively. The preference shock follows standard normal distribution. We also include normally distributed preference heterogeneity as in Equation 5 and the standard deviations of the heterogeneity are set at 0.5. For the purpose of evaluating the role of exclusion restrictions, we vary the specifications of search costs for these datasets:

- In the first dataset, the search cost only has a constant term with heterogeneity. The search cost constant is set at 2. The heterogeneity of the search cost constant is normally distributed with a standard deviation of 0.5. In other words, there are no exclusion restriction variables in z_{ij} and the separation between the search cost and the utility only relies on the nonlinearity relationship between the two (Equation 4). More importantly, when making draws for the heterogeneity terms, we introduce high correlation between the constant terms of the utility and the search cost, with a correlation coefficient of -0.9. By doing so, we try to introduce correlation between the utility and the search cost that the nonlinearity alone cannot easily eliminate.

- In the second dataset, we has two additional covariates in the search cost besides the constant term, namely time-constraint of the consumer and slot position of a product. In the search cost function, time-constraint and slot position have true coefficients of 1 and -1. The heterogeneity of these two covariates is normally distributed with standard deviations both being set at 0.3. The two additional covariates do not enter utility function. Accordingly, they can be viewed as exclusion restriction variables. The constant terms of utility and search cost remain correlated.

We then create another two datasets with the same setting as the first two but with 400 consumers in each dataset. Using the estimation approach proposed, Table 1 shows the results across these four simulated datasets. From the Table, we have the following observations:

- In the first dataset (200 consumers), there are no exclusion restriction variables and the utility and the search cost are correlated. The estimates of the constant terms

are insignificant, which implies that they cannot be identified. More data in the third dataset (400 consumers) do not help the identification as evidenced by the insignificant estimates.

- In the second dataset (200 consumers), there are exclusion restrictions and the constants of the utility and the search cost remain correlated. We are able to recover the true parameters. In comparison to the first and the third datasets, the exclusion restrictions help to eliminate the correlation between the utility and search cost because the second dataset has additional covariates which are orthogonal to the utility. More data in the fourth dataset (400 consumers) further enhance estimation efficiency by decreasing the estimates' standard errors.

In conclusion, the simulations demonstrate that the exclusion restriction variables are able to separate the search cost and the utility. Especially, when the search cost and the utility are correlated, the nonlinearity embedded in the model may not be sufficient to separately identify the search cost from the utility.

[Insert Table 1 About Here]

5 Application: Click-stream Data of Hotel Booking

To exemplify its applicability, we apply the proposed model and estimation approach to a click-stream dataset of hotel bookings provided by a major website of travel products.

5.1 The Website

When visiting this website, the consumer first specifies the product of interest, such as the location, check-in and check-out dates, etc. The website shows the consumer a list of hotels that satisfies the criteria. If the consumer is satisfied with one hotel, she completes the purchase by booking the hotel through the website.

The list of hotels can be very long. The website displays up to 25 hotels per page. The consumer can then choose to explore the next 25 hotels on the list by turning to the next

webpage. Even on a given webpage, however, the computer screen size can make viewing all 25 hotels at once difficult. The consumer can view about four hotels on the list with a reasonably high-resolution computer screen (e.g. 1,920 by 1,200). To view the remaining hotels on the list, the consumer has to scroll up or down the page. The list contains summary information on average daily price, star ratings, and consumer review ratings for each hotel. To obtain more detailed information, such as total price with fees and taxes, detailed reviews, and amenities, the consumer has to click through the hotel's link.

The list of hotels is sorted according to a default ranking by the website. The website's management team explained to us that the default ranking is based on the numbers of bookings during the previous period. However, this rule is not disclosed to consumers. The default list is named "[Website] Picks," and the ranking rule is vaguely described in the FAQ section as "the summary from the most affordable price, highest guest rating, highest star rating, and the hotel nearest to the airport, to the expensive price, lowest guest rating, lowest star rating, and the hotel farthest to the airport."

The consumer can refine the default search results using alternative sorting and/or filtering methods (e.g., sort by prices, filter by star ratings, etc.). After the refinement, if two hotels have the same level of the attribute used for the refinement (e.g., when being sorted by star ratings, both hotels have a four-star rating), they will be ranked according to the default ranking algorithm.

5.2 Data

5.2.1 Click-stream Data

The dataset contains 495 individual consumers' click-stream data between October 1 and October 15, 2009, for their hotel search and purchase activities. Each consumer searched hotels in one of the four cities: Budapest, Cancun, Manhattan, and Paris; and each booked one hotel after the search (i.e., 495 purchases in total). In this study, we apply our model to consumers who made purchases on the website and also assume that all consumers make

purchases after search in our counterfactual analyses. Note that our model is general enough to allow the analysis of data with observations on purchases of an outside good. However, we do not have information in our current dataset regarding the nature of outside goods, which may include no travel, booking hotels from other websites or local travel agencies, etc. While this is a limitation of our study, focusing on consumers who have made purchases does have its own managerial importance. By understanding the search behavior of these customers, the firm can enhance the shopping experience and thus consumers' overall satisfaction through a better default ranking design and the communication of the ranking policy. Many companies, especially those in the service industry, e.g., our data providing website, are contriving to improve the shopping experience and satisfaction of their own customers.

By website design, consumers need to click through the hotel's link for detailed information before a purchase. Table 2 and Figure 1 present the summary statistics and histogram of consumer click-through activities, respectively. These 495 consumers made a total of 1,140 click-throughs, with an average of 2.30 click-throughs per consumer. However, there was a large variation across individuals.

[Insert Table 2 About Here]

[Insert Figure 1 About Here]

On average, consumers book their hotels about 4 weeks in advance. However, the lapse between the day of the search and the check-in date varies greatly across consumers. Table 3 shows the summary statistics.

[Insert Table 3 About Here]

Consumers' refinement activities also exhibited great diversity. The number of refinement activities among consumers ranged from 0 to 6. The diversity of refinement methods used indicates that the consumers may be heterogeneous in their preferences about hotel attributes. The top seven sorting/filtering methods accounted for 86% of all refinement activities: (1) sort by price ascendingly, (2) sort by consumer review rating descendingly, (3)

filter out hotels below 4-star, (4) filter out hotels below 3-star, (5) filter out hotels below 3-star and sort by price ascendingly, (6) filter out hotels below 5-star, and (7) filter out hotels below 4-star and sort by price ascendingly. We group the default list and all the other less-used refinement methods as the eighth option, "no refinement".

There are 282 consumers who used at least one of the top seven refinement methods, with an average of 1.74 refinement activity per person (492 refinements in total) and a standard deviation of 1.09. Table 4 and Figure 2 present the distribution of the refinement activities among these 282 consumers. Furthermore, recall that we define a click-through as a search in the model. In the data, 90.40% of the top seven refinement activities were followed by at least one click-through. Figure 3 shows the histogram of click-throughs after each refinement activity. In total, these 282 consumers who had refinement activities accounted for 759 click-throughs, 489 of which were made after refinement activities.

[Insert Table 4 About Here]

[Insert Figure 2 About Here]

[Insert Figure 3 About Here]

5.2.2 Hotels

On the supply side, there was a total of 1,961 hotels. Note that depending on the city searched by each consumer, each consumer was only shown a city-specific subset of these 1961 hotels. We observed basic hotel attributes, including daily price, star rating, consumer rating, distance to city center, whether a hotel is affiliated with a hotel chain, and whether a "promotion" flag is displayed beside the hotel link. Table 5 reports summary statistics of hotel attributes overall and among clicked hotels.

[Insert Table 5 About Here]

5.3 Application of the Model and Estimation

In this section we describe the application of the model and estimation approach to this dataset.

5.3.1 Utility

The utility is specified as:

$$u_{ij} \quad = \quad x'_{ij} \cdot (\alpha + \xi_i \odot \sigma_u) + \nu_{ij}$$

where x_{ij} is a vector of hotel attributes, including the hotel attributes shown in Table 5 and city intercepts.

Because every consumer made one purchase, no one chose the outside option in the data (i.e., not purchasing from the focal website). Accordingly, we normalize the mean sensitivity of price as -1 for identification purpose. Normalizing the mean price sensitivity to -1 allows us to scale other parameters accordingly and interpret them against one dollar.

One concern is that hotel price may be endogenous due to some unobserved hotel attribute included in the preference shock ν_{ij} (e.g., price is positively correlated with unobserved hotel quality contained in ν_{ij}). Correspondingly, we treat hotel price as endogenous and use instrument variables. The instruments are chosen in the spirit of Berry et al. (1995) and Hortacsu and Syverson (2004) for estimating unobserved quality. We constructed the instruments so that they affect hotel pricing decisions but are independent of the unobserved hotel quality. In particular, we use average prices of the same market (excluding the focal hotel) and own hotel non-price attributes. To incorporate the instruments into our nonlinear estimation, we adopt the control function approach proposed in Petrin and Train (2010).

5.3.2 Heterogeneous Search Cost

To capture the heterogeneity in search cost, we specify search cost as[6]

$$\begin{aligned} c_{ij} &= c_i(TimeConstraint_i, Slot_j) \\ &= \exp(\gamma_{i0} + \gamma_{i1} TimeConstraint_i + \gamma_{i2} Slot_j) \\ &= \exp(\gamma_0 + \zeta_{i0}\sigma_{0c} + (\gamma_1 + \zeta_{i1}\sigma_{1c}) TimeConstraint_i + (\gamma_2 + \zeta_{i2}\sigma_{2c}) Slot_j) \end{aligned} \tag{11}$$

where $TimeConstraint_i$ is the number of days between consumer i's search and her check-in. $Slot_j$ is the slot position of the j-th search. The exponential operator is to assure that the costs are positive. γ_0, γ_1, and γ_2 are mean levels of cost coefficients. $\zeta_{i0}\sigma_{0c}$, $\zeta_{i1}\sigma_{1c}$, and $\zeta_{i2}\sigma_{2c}$ are the deviations from the mean levels and capture the heterogeneity in search cost. $TimeConstraint_i$ and $Slot_j$ do not enter one's utility function. They serve as the exclusion restrictions and help to separate the search cost from the utility.

5.3.3 Refinement

For the j-th search, the consumer needs to decide which slot on the list to search. In particular, the distribution of product attributes x_{ij} may depend on the slot. The default order of the hotel list has no obvious ordering on hotels' attributes. Consumers do not know that the default ranking is based on the booking frequencies of hotels.[7] As a result, from the perspective of the consumer, it can be considered that x_{ij} is randomly drawn from the attributes distribution independent of $Slot_j$.

However, consumers have the option of refining the search results using sorting and filtering. Refinement will affect the distribution of x_{ij} on the given slot position $Slot_j$ in the following ways:

[6] We also consider two alternative search cost specifications in Section 5.4.3 and confirm the current specification is appropriate.

[7] In Section 5.4, we implement a series of robustness checks for this assumption and confirm its validity.

- **The effect of sorting on the distribution of x_{ij}.** If the consumer sorts the hotels based on some attribute such as price, the sorted attribute becomes an ordered statistic. For example, if the hotels are sorted by price ascendingly, then the consumer knows that hotels on a lower slot position on average have higher prices than the one on the first slot.[8] Since other attributes are likely to be correlated with price, the sorting will also have an impact on the levels of those attributes conditioned on the slot position.

- **The effect of filtering on the distribution of x_{ij}.** Filtering on a specific attribute eliminates hotels that do not meet the criterion. As a result, the filtering changes the attribute distribution of the listed hotels. For example, if the consumer uses the filter to show only five-star hotels, then the distribution of star ratings is truncated below five-star. Since star ratings and other attributes (e.g., price) are correlated, such a filtering also affects the levels of other attributes.

To accommodate these effects of refinement, we allow $P_j(x_{ij})$, the attributes distribution of the j-th search, to be sorting/filtering-specific. For a given slot position $Slot_j$ and a given sorting/filtering method k ($k = 1, 2, ..., 7, 8$), $P_j(x_{ij}) = P^k(x_{ij}|Slot_j)$ is the attributes distribution of the j-th search.

5.3.4 Single-level Discrete Choice vs. Multiple-level Discrete Choice

In this application, we define a search as exploring a slot position using a refinement method. This definition implies that the consumer makes discrete choices among different slot positions across different refinement methods. Alternatively, in a multiple-level discrete choice model, the consumer first chooses the refinement method, then decides the slots to search conditioned on the refinement decision. We choose the current single-level specification due to the following reasons:

[8] When a consumer sorts the hotels based on a particular attribute, e.g., price, the hotels may not be completely sorted according to that attribute. This is because the website sometimes features certain promotional items on the list and the refinement does not apply to those featured items. As a result, even though on average the products are sorted according to the attribute of interest, there may still be some uncertainty involved.

1. The value of the random term in the utility function stays the same for a particular product across different refinement methods. This implies that a given product has the same utility level across refinement methods – a fairly reasonable specification as refinement during the search process is unlikely to affect the consumption utility of a product. Furthermore, under the rational expectation framework, a consumer forms her expectation about the attribute levels of each slot position across refinement methods; for the same slot position the search cost remains constant across refinement methods. There is no additional randomness at the refinement level. As a result, the current single-level specification and the multiple-level specification lead to equivalent likelihood, because at the refinement level the comparison is among deterministic expectations.

2. It is unclear whether the consumer decides on the slot position or the refinement first. It is possible that a consumer with high search cost decides to search only top slots, then she decides on the refinement to assure that the distribution of attributes on the top slots leads to higher expected utility levels. Since the data cannot distinguish the sequence of these decisions, a single-level model becomes a more natural choice.

3. In the classical discrete choice literature (e.g., consumer brand/quantity decision), it is common to use either multiple-level or single-level models. For multiple-level models, a consumer first chooses the brand and then determines the quantity depending on the brand choice (e.g., Chiang (1991); Chintagunta (1993); Arora et al. (1998); Nair et al. (2005)). For single-level models, a consumer treats each brand-quantity combination as a choice alternative (e.g., Guadagni and Little (1983); Chintagunta (1992, 1998)).[9] This application is somewhat similar to the single-level models.

5.4 Results and Robustness Tests

In this section we report the results of the estimation and fit information of the application.

[9]For more complete reviews, see Allenby et al. (2004) and Chintagunta and Nair (2011).

5.4.1 Parameter Estimates

Table 6 reports the parameter estimates. Besides city intercepts, consumer ratings have on average the highest impact on consumer utility. If a hotel has a consumer rating between 4 and 4.5, all other things being equal, the hotel may set its daily price $78.11 higher than hotels with a rating lower than 4. If the rating is above 4.5, the premium increases to $106.56. Other significant factors that affect utility are star rating, promotion, and chain affiliation.

Search cost is significant. It has important implications for consumer search behavior. Hotels that appear lower in the ranking of slots have lower chances of being searched. Placing hotels with high expected utility levels in more prominent positions may reduce the total number of searches and therefore the total search cost. Hence, the existence of search cost makes refinement especially beneficial for consumers.

We also find that consumers demonstrate considerable variation in preference and search cost. For example, although the mean level of the baseline search cost is around $25.28 (=exp(3.23)), there are significant variations, particularly due to people's time constraints. If a consumer searches hotels 30 days in advance (the population average), the search cost drops more than 2/3 from the value when the search happens on the same day of the check-in (exp(-0.04·30) vs. exp(-0.04)). Similarly, people demonstrate considerable heterogeneity in their sensitivities for product attributes. As a result, consumers may use alternative refinement methods that prioritize more important attributes. We will further explore the ramification of heterogeneity on market structure in the policy simulation section.

[Insert Table 6 About Here]

5.4.2 Model Validation

To examine the fit of the model, we consider three tests using a holdout dataset. We randomly select 100 consumers from the 495 as a holdout sample (about 20% of the full sample). We then estimate the model using only 395 consumers.

We begin by calculating the hit rates of hotel search and purchase. For each individual consumer, we use the model estimates to generate 100 sets of preference and search cost parameter values, as well as 100 random utility shocks per hotel per consumer (v_{ij}). Conditioned on the observed hotel attributes levels, prices, and slot positions, for each set of parameter values and consumer-hotel random shocks, we infer which hotel is searched and, conditioned on the searches, which hotel is booked by consumer i. We repeat the exercise using all parameter draws and random utility shocks, and then calculate the hit rate. We find the hit rates are 0.82 and 0.69 for search and purchase, respectively, suggesting our model captures the search behavior well.

Consumers have heterogenous sensitivities for each of the hotel attributes, which is one reason why they use different refinement methods. To the extent that consumers' choices of refinement methods reflect their heterogeneity, we consider a another test to examine the model's ability to recover the heterogeneity. Using a similar approach, we have 0.73 as the hit rate for the eight sorting/filtering methods (the top seven plus "no-refinement").

As for in-sample fit, the corresponding hit rates are 0.83, 0.70, and 0.76, respectively. Overall the model fits well.

5.4.3 Robustness Tests

In this section, we consider several robustness tests pertaining to the current specification of the model.

Alternative Information Structure By default, the search results of hotels are ranked according to the frequencies of purchases. However, since the website does not publicize information about this default ranking rule and even obscures it, we assume that consumers do not know the rule. This assumption implies that when the consumers view the default list, they treat the hotel attributes independent from slot positions. To be consistent with this assumption in the estimation, when a consumer faces the default ranking, for each slot

we randomly draw the attributes from a joint distribution that is independent of the slot position and obtained from the data.

It is possible, however, that consumers know the default rule during the search. In particular, they may infer that top hotels on the default list are more popular (higher frequencies of bookings). This alternative assumption implies that, under the default ranking, consumers know that more preferable attribute levels are more likely to be observed at top positions than at inferior positions.

To evaluate this alternative assumption, we consider two tests:

1. We re-estimate the model under the alternative assumption, i.e., consumers understand the default ranking rule. In particular, when a consumer faces the default ranking, instead of drawing attributes from a distribution that is independent of slot positions, we draw them from slot-specific distributions, obtained as the empirical distributions from the data. The out-of-sample fit deteriorates as measured by the hit rates as in Section 5.4.2. The measures change from 0.82, 0.69, and 0.73 to 0.65, 0.52, and 0.70. We take these as evidence that the original assumption (consumers do not know the rule) is more appropriate for the data.

2. About 29% of consumers in the data (146 in total) can be identified as "frequent users" since they auto-logged into their accounts upon their arrivals at the website. It is reasonable to expect that, if some consumers understand the default ranking rule, it is more likely to be these 29% frequent users. We use these "frequent users" to re-estimate the model under the original assumption and the alternative one. We calculate the in-sample fit using the hit rates.[10] The hit rates under the original assumption are 0.80, 0.66, and 0.70 for these frequent users. In comparison, the measures become 0.76, 0.64, and 0.65 under the alternative assumption. This result implies that even for those who

[10] We choose to use in-sample fit instead of out-of-sample fit because of the much smaller size of the sample (only 146 frequent users).

are more likely to understand the default ranking rule, the original assumption seems more appropriate.

Alternative Search Cost Specifications The search cost is specified as

$$\begin{aligned} c_{ij} &= c_i(TimeConstraint_i, Slot_j) \\ &= \exp(\gamma_{i0} + \gamma_{i1} TimeConstraint_i + \gamma_{i2} Slot_j) \end{aligned} \tag{12}$$

We also consider two alternative specifications regarding search cost:

1. The search cost of a slot is determined by the page number on which it is located. Slots on the same webpage share the same search cost. Under this specification, we have

$$\begin{aligned} c_{ij} &= c_i(TimeConstraint_i, Page_j) \\ &= \exp(\gamma_{i0} + \gamma_{i1} TimeConstraint_i + \gamma_{i2} Page_j) \end{aligned}$$

   where Page_j is the webpage number where $Slot_j$ is located. For example, slot 51 to slot 75 have the same cost level because they are on the same webpage (25 hotels per page).

2. The search cost of a slot is determined by both the slot position and the number of pages. In particular,

$$\begin{aligned} c_{ij} &= c_i(TimeConstraint_i, Slot_j, Page_j) \\ &= \exp(\gamma_{i0} + \gamma_{i1} TimeConstraint_i + \gamma_{i2} Slot_j + \gamma_{i3} Page_j) \end{aligned}$$

   where $1 \leq Slot_j \leq 25$ is the slot position on a specific webpage, and Page_j is the webpage number. For example, for slot 51, $\text{Page}_j = 3$ and $Slot_j = 1$ (i.e., the first slot on page 3.).

Under these two alternative search cost specifications, the utility estimates are essentially the same but the model fit deteriorates as measured by out-of-sample fit (0.82, 0.69, 0.73 versus 0.81, 0.62, 0.70; 0.82, 0.69, 0.73 versus 0.79, 0.65, 0.69). More importantly, the coefficients of Page_j are insignificant in both alternative specifications, potentially due to the sparse observations of page-turning among consumers.

Cost of Refinement We next consider the implications of refinement on search cost.

First, we assume that search cost is not specific to refinement method, and for a given slot position, the search cost stays fixed across refinement methods. However, it is possible that the search cost may depend on the refinement method used. To provide more support for our assumption, we re-estimate the model under two new settings using the cost function in Equation 12. In the first new setting, we constrain the refinement methods either to sorting alone or to others, where the latter includes "no refinement" and refinements involving filtering. In the second new setting, we constrain the refinement methods either to filtering alone or to others, where the latter includes "no refinement" and refinements involving sorting. The focal website lists filtering options farther down the page than sorting options. Accordingly, if the search cost varies across refinement methods, we would expect filtering options to have higher or at least different cost levels than sorting options. Consequently, we would expect the estimation results from those two new settings be different from each other and from the current setting in the paper. However, we find that the results are statistically equivalent across the three alternative settings, with the current one having the best fit. We consider such an observation as evidence that the search cost of a given slot position is fixed across refinement methods.

Second, the action of switching refinement methods may incur additional costs beyond search. To explore such a conjecture, we consider two robustness checks.

1. Divide the 282 consumers who used refinement into two groups: (1) those who used one refinement method (165 consumers); and (2) those who used at least two refinement methods (117 consumers). We then re-estimate the model using both samples with

the cost specification in Equation 12. If switching refinement methods is costly, the current cost function (Equation 12) is mis-specified. As a result, since the two groups on average have different numbers of refinement activities, the new estimates of the two groups are expected to be different from each other and from those presented in Table 6. However, we find that the estimates of both samples are not statistically different from each other and from the current estimates.

2. Divide the 282 consumers who used refinement into two groups, (1) those consumers who made at most one search after each refinement activity (136 consumers), and (2) those consumers who made more than one search after each refinement activity (146 consumers). Again, we re-estimate the model using both samples. Similarly, if switching refinement methods has additional costs, the model suffers mis-specification. In that case, we would expect the estimates across the two groups and those shown in Table 6 to be different. On the contrary, the estimates of the two groups are statistically equivalent to one another and to those reported estimates.

Based on these robustness checks, we conclude that switching refinement methods has little effect on cost.

5.5 Managerial Implications

5.5.1 Refinement and Consumer Welfare

The Effect of Refinement on Consumer Welfare High search cost limits consumers' searches and may force them to choose options with lower utilities. As discussed earlier, giving consumers the ability to refine search results along the dimensions that matter most to them may reduce their search costs. To empirically investigate this insight, we simulate the searching and booking outcomes of all consumers without and with the refinement ability. When the sorting/filtering options are removed, the average number of searches across consumers is 1.69. With the sorting/filter options, however, each consumer on average makes

2.27 searches, a 34% increase.[11] The average utility for hotels booked increases by 18% when refinement tools are available.

Although the final utility for the booked hotel becomes higher with refinement, it is still unclear what impact refinement tools have on the overall welfare of consumers. If the number of searches increases disproportionately to the utility improvement with refinement, the accumulated search costs may well outweigh the benefits, hence lowering the overall consumer welfare. To evaluate the overall consumer welfare, we further compute the net surplus of search as measured by the final utility of the booked hotel net the total search cost. Surprisingly, we find that on average the net surplus decreases by 1.8% with refinement, and the 95% confidence interval of the welfare loss is (-3.4%, -0.2%). To better understand this seemingly counter-intuitive result, recall that the default ranking of hotels is based on booking frequencies, which to some extent already reflects the average utility levels of these hotels among the population. Consequently, even without refinement tools, the baseline level of consumer welfare is fairly high if consumers make decisions according to the default ranking. As a result, the main reason for the welfare reduction with refinement is that consumers do not understand the default ranking rule and disproportionately made more searches.

To explore this insight, we consider an additional simulation. We again simulate the searching and booking outcomes of all consumers with and without the refinement ability. However, in this simulation we assume that consumers are educated about the default ranking rule. Under the default ranking of hotels, attributes are drawn from slot-specific distributions. Through this simulation, we find that the average number of searches with the availability of refinement is 2.07, compared to 1.60 without refinement tools. These numbers are smaller than those made by consumers uninformed about the default ranking rule (2.27 and 1.69, respectively). We also find that net welfare surplus increases by 1.3% with the refinement, compared to the drop of 1.8% with uninformed consumers. The 95%

[11] This is consistent with the data where we observe on average 2.30 searches per consumer.

confidence interval of the welfare improvement is (0.9%, 4.1%). This result is consistent with our conjecture, i.e., consumers who are uninformed of the default ranking rule engage in disproportionately more searches, leading to the deterioration in net welfare surplus. In contrast, refinement tools improve the net welfare surplus when consumers understand the default ranking rule.

While the increase in welfare seems small, the overall effect may be considerable when we take into account the size of consumer population at the website. Furthermore, the firm can revise its website with little investment. A simple clarification to consumers about the default ranking rule will improve their satisfaction.

Alternative Default Ranking In addition to educating consumers about the default ranking rule, it is possible to further enhance net consumer surplus by providing an alternative default ranking scheme using additional information. In particular, Ghose et al. (2012) and Ghose et al. (2014) show that a website can improve consumer welfare by directly ranking products by consumer utility levels.

Recall that if two hotels have equivalent ranking after the refinement, they will further be ranked according to the default ranking, which is based on booking frequencies. However, booking frequencies may not perfectly reflect utility levels. First, under the current default rule, while a higher booking frequency can secure a more prominent slot, a better slot can also enhance the booking frequency. Such a self-fulfilling effect may deteriorate the default ranking's ability to approximate the ranking of actual utility levels. Second, since booking frequencies are calculated based on historical data, it will take some time for the default ranking to reflect any utility changes. For example, when a hotel decreases its price during a promotion, it may achieve a high booking frequency. When the promotion ends, the utility level will fall due to the high regular price. However, this will not be reflected immediately in the default ranking.

Accordingly, we propose using the inferred utility levels based on the model to rank the hotels directly. To investigate the effect of this alternative ranking method on consumer welfare, we consider the following policy simulation:

1. Conditional on observed hotel attribute levels and model estimates, we obtain the mean utility of each hotel.

2. Instead of using the observed default ranking of hotels, we rank the hotels based on the imputed mean utilities in Step 1. In particular, after sorting/filtering, if two hotels have the same implied ranking based on the refinement method, the ranking will further be determined by their mean utilities.

3. We assume that consumers know this new default ranking rule. Then for each consumer, we generate 100 sets of parameter values as well as 100 random utility shocks per hotel.

   (a) Conditioned on the observed hotel attributes levels, prices, and new positions based on the mean utility ranking, we can infer which hotel this consumer will search for a given set of parameter values and hotel-consumer random shocks. We can then compute the utility of the booked hotel and the net welfare of this consumer.

   (b) To compute the *expected* utility of the booked hotel and the net welfare of this consumer, we need to integrate the distributions of preference, search cost, and random shocks by repeating Step 3(a), using all 100 sets of parameter draws and random utility shocks, and then calculating the average.

4. We repeat Step 3 for all consumers and then aggregate the results to calculate the total net welfare and total utility of booked hotels.

In comparison to the observed default ranking of the website, the utility of the booked hotels increases by 2% with a 95% confidence interval of (0.7%, 4.6%). The total net welfare of

consumers increases by 2.9% with a 95% confidence interval of (0.9%, 5.2%). This simulation result implies that the combination of refinement tools and the new ranking by mean utilities would enhance the overall net welfare surplus by about 1.6%.[12] We compare the average number of searches under the alternative ranking method and the current default ranking method. Also, we assume that consumers are informed about the ranking rules in both cases. We find that under the new ranking method, the number of searches is 2.00, lower than the 2.07 searches under the current default ranking method with informed consumers. In other words, improvement in the net welfare comes from both the enhancement of the utility of the booked hotels and the decrease in the number of searches.

5.5.2 Refinement and Market Structure

The ability to refining search results may affect the market structure. Consumers face the same slot ranking under the default list. It may be too costly for a consumer to reach preferable hotels if they are ranked low on the default list. It is possible that most consumers without refinement tools will be limited to the top-ranked hotels on the default list due to search cost, even though they would have chosen differently otherwise. Only consumers with relatively lower cost may search farther down the list. Consequently, the top-ranked hotels on the default list tend to have higher market shares. In comparison, consumers with sorting and filtering capabilities will use different methods because they have heterogeneous preferences. The choices are no longer limited to the top hotels on the default list. Thus the market becomes more competitive.

To explore the impact of refinement on market structure, we start by calculating the Herfindahl-Hirschman Index of search shares of hotels under the current market condition: heterogenous consumers with refinement options.[13] Herfindahl-Hirschman Index (HHI) is a

[12]The percentage 1.6% is computed as (2.9%-1.3%). The benchmark 1.3% is the welfare improvement when consumers are informed about the current default ranking rule (the counterfactual considered above).

[13]We choose to use search shares instead of purchase shares of hotels. The reason is that there are 1,961 hotels but only 495 purchases. The purchased hotels may not be representative, which makes the calibrated purchase shares less meaningful.

measure of the intensity of market competition, defined as

$$HHI = \sum_{j=1}^{H} s_j^2$$

where s_j is the market share of firm j. According to the U.S. Department of Justice and the Federal Trade Commission, an HHI index between 0.15 to 0.25 indicates moderate competition. An HHI index above 0.25 implies a highly concentrated market structure that lacks competition.[14] We calculate search shares using the same method as in Section 5.4.2. Under the current market condition, the HHI takes the value of 0.16, indicating moderate competition in the market.

Next, we remove the refinement options so that all consumers face the same default hotel list. Under this new market condition, the HHI increases to 0.29, showing a high level of market concentration among the top hotels on the default list. We further remove the heterogeneity of preference and search cost among consumers. In this case, the HHI increases by another 24%, reaching 0.36. In conclusion, the refinement combined with the heterogeneity of consumers makes the market less concentrated and more competitive.

6 Conclusion

This paper proposes a structural model of consumer optimal sequential search. Using click-stream data of individual online purchase and search activities, we show that the model can be identified. The identification relies on the exclusion restrictions separating the search cost and the utility. Such exclusion restriction variables are easier to obtain from click-stream data. We are able to estimate the preferences and search costs of heterogeneous consumers, providing insights about consumer decisions in face of uncertainty about product attribute levels.

Furthermore, the impact of search technology is of great interest to both industry and academia. In particular, the ability of consumers to sort and filter search results has sub-

[14] Horizontal Merger Guidelines, the U.S. Department of Justice and the Federal Trade Commission, 2010.

stantial effects on consumer and firm behavior. In our model, consumer decisions of refining search results can also be incorporated into the framework.

Our modeling approach has a few important features. First, the identification strategy and the corresponding estimation approach enable the model to be applied to other online search contexts where consumer-level click-stream data are commonly available. As a result, the model has broad applicability. Second, the model explicitly treats consumer search as a utility maximization process. The model is consistent with classical optimal information search theory and has a solid theoretical foundation. Third, although it may be unrealistic, many previous studies on consumer choices assume that consumers have perfect knowledge about product attributes for the sake of tractability. Instead, our model allows uncertainty to be resolved during the search. More importantly, a consumer's refinement decision will affect how uncertainty being resolved by changing the distribution of product attributes across slot positions on web pages. As a result, decisions of search and refinement are coherently integrated into the utility optimization framework.

We apply the model and estimation to a travel website's click-stream dataset. The application of the model shows a decent out-of-sample fit. In particular, it has the ability to recover the pattern of consumer heterogeneity. Conditioned on the estimates, we consider several policy simulations. First, we find that, with the aid of refinement tools, consumers make 34% more searches on average and are able to obtain 18% higher utilities from the products they choose. Second, although the utility levels of the purchased products increase, the overall welfare surplus may drop for consumers. The welfare reduction occurs when consumers do not understand the website's default ranking rule and disproportionately make excessive searches using refinement tools. The default ranking is based on the booking frequencies of hotels, which to some extent already reflects the qualities of the hotels. As a result, the baseline level of consumer welfare is fairly high even without refinement tools. To address such deterioration in net welfare, we show that simply educating consumers about this rule will improve consumer welfare. We also suggest a new ranking method that can

further enhance consumer welfare. This new rule uses imputed mean utilities of products to determine ranking. The new default ranking method has the ability of improving both the utility of purchased product and the net welfare. Third, we also find that refinement tools make the market less concentrated, because they help heterogeneous consumers find hotels that match their preferences better. Such matches would be too costly without refinement tools.

Several extensions to the current model are possible. First, although consumers in our model do not know product attributes before the search and use search to resolve uncertainty, we assume consumers know the distribution of attributes. This assumption is reasonable in the current context as attribute levels of hotels across time are relatively stable and consumers are likely to be familiar with the marketplace. However, in contexts where consumers face some unfamiliar product category, the distribution may also be unknown and consumers need to update their beliefs about the distribution based on every round of search. Adam (2001) proposes a theoretical optimal search model where the agent learns the profits distribution of alternative options during her search process. Koulayev (2013b) estimates a model where consumers update their Dirichlet prior beliefs on price during the search process. De los Santos et al. (2013) consider a model where consumers have uncertainty about the overall utility distribution. During the search process, they update their Dirichlet prior beliefs under a Bayesian learning framework. In short, integrating learning into the model will certainly enhance our understanding of the consumer search process.

Second, the website can be interpreted as a platform of a two-sided market that facilitates interactions between sellers and consumers (Yao and Mela (2008, 2011)). The current model focuses on the demand side. Extending the model to include the supply side will further enrich our insights into such markets and will enable additional policy simulations such as the strategic interactions among the sellers.

Finally, an examination of how consumers and firms adapt to the advance of search technology in the long-term will be a fruitful avenue for future research. For example, the

advance of search technology enables consumers to search more extensively for lower prices, which intensifies price competition among firms. Ellison and Ellison (2009) show that, to minimize damages, firms may start to engage in information obfuscation, making obtaining their product information from the search engine more difficult for consumers. Overall, we hope this paper will inspire future research on consumer online search.

Table 1: Monte Carlo Simulations

True Values of Parameters	Dataset 1: Without Exclusion Restrictions, Correlated Utility and Search Cost (N=200)	Dataset 2: With Exclusion Restrictions, Correlated Utility and Search Cost (N=200)	Dataset 3: Without Exclusion Restrictions, Correlated Utility and Search Cost (N=400)	Dataset 4: With Exclusion Restrictions, Correlated Utility and Search Cost (N=400)
Utility Constant=5	3.07 (3.33)	**4.77 (0.35)**	2.77 (2.48)	**4.86 (0.30)**
Utility Constant Heterogeneity=0.5	0.82 (0.50)	**0.44 (0.11)**	0.62 (0.40)	**0.46 (0.05)**
Price=-2	**-2.80 (0.88)**	**-2.19 (0.49)**	**-2.50 (0.63)**	**-2.20 (0.30)**
Price Heterogeneity=0.5	**0.65 (0.15)**	**0.41 (0.09)**	**0.65 (0.15)**	**0.57 (0.07)**
Quality=2	**2.71 (0.99)**	**2.30 (0.77)**	**2.51 (0.69)**	**2.23 (0.70)**
Quality Heterogeneity=0.5	**0.41 (0.14)**	**0.40 (0.11)**	**0.43 (0.10)**	**0.44 (0.06)**
Search Cost Constant=2	2.80 (1.70)	**2.19 (0.39)**	3.10 (2.73)	**2.12 (0.29)**
Search Cost Constant Heterogeneity=0.5	0.70 (0.40)	**0.42 (0.08)**	0.61 (0.30)	**0.56 (0.05)**
Time Constraint=1	-	**0.89 (0.20)**	-	**0.90 (0.17)**
Time Constraint Heterogeneity=0.3	-	**0.49 (0.15)**	-	**0.39 (0.10)**
Slot Position=-1	-	**-1.22 (0.32)**	-	**-1.12 (0.21)**
Slot Position Heterogeneity=0.3	-	**0.45 (0.13)**	-	**0.40 (0.09)**

Note: Bold fonts indicate the estimates being significant at 95% level.

Table 2: Summary Statistics of Consumers' Click-throughs

	Mean	Std. Dev.	Min.	Max.
Click-throughs per Consumer (495 consumers)	2.30	2.50	1	22
– Budapest (237 consumers)	2.08	2.28	1	22
– Cancun (74 consumers)	1.91	1.65	1	10
– Manhattan (97 consumers)	2.87	3.09	1	18
– Paris (87 consumers)	2.61	2.83	1	15

Table 3: Summary Statistics of Consumers Lapse between the Search and the Check-in

	Mean	Std. Dev.	Min.	Max.
Days between the Search and the Checkin (495 consumers)	29.91	42.51	0	327
– Budapest (237 consumers)	28.69	46.37	0	327
– Cancun (74 consumers)	35.62	41.35	0	200
– Manhattan (97 consumers)	27.10	31.53	0	168
– Paris (87 consumers)	31.47	43.29	0	242

Table 4: Summary Statistics of Top Seven Refinement Methods

	Mean	Std. Dev.	Min.	Max.
Refinement per Consumer (282 consumers)	1.74	1.09	1	6
– Budapest (131 consumers)	1.65	1.03	1	6
– Cancun (34 consumers)	1.65	0.98	1	4
– Manhattan (62 consumers)	1.73	1.18	1	5
– Paris (55 consumers)	2.04	1.17	1	6

Table 5: Summary Statistics of the Hotels

	Mean	Std.Dev.	Mean of Clicked Hotels	Std.Dev. of Clicked Hotels
Budapest (276 hotels)				
Average Daily Price ($)	82.64	31.43	78.90	28.08
Star Rating	3.37	1.47	3.70	1.16
Consumer Rating	2.25	2.15	2.89	2.06
Distance to City Center (km)	3.01	3.54	2.53	2.89
Hotel Chain	0.31	0.46	0.40	0.49
Promotion Flag	0.30	0.46	0.49	0.50
Cancun (106 hotels)				
Average Daily Price ($)	144.54	53.89	129.88	43.14
Star Rating	3.52	1.04	3.58	0.67
Consumer Rating	3.55	1.64	3.67	1.41
Distance to City Center (km)	8.83	3.61	9.13	2.75
Hotel Chain	0.40	0.49	0.38	0.49
Promotion Flag	0.70	0.46	0.86	0.35
Manhattan (487 hotels)				
Average Daily Price ($)	252.65	96.65	233.84	76.29
Star Rating	3.21	0.90	3.09	0.78
Consumer Rating	3.54	1.45	3.59	1.31
Distance to City Center (km)	2.15	1.74	2.01	1.62
Hotel Chain	0.43	0.50	0.39	0.49
Promotion Flag	0.31	0.46	0.46	0.50
Paris (1092 hotels)				
Average Daily Price ($)	155.91	48.45	158.09	45.93
Star Rating	3.02	1.10	3.04	1.14
Consumer Rating	1.92	1.98	2.26	1.99
Distance to City Center (km)	4.32	5.16	4.33	4.65
Hotel Chain	0.47	0.50	0.40	0.49
Promotion Flag	0.48	0.50	0.51	0.50

Table 6: Model Estimates*

	Mean Parameters (Std. Err.)	Heterogeneity (Std. Err.)
Search Cost		
Constant	**3.23 (0.53)**	**1.30 (0.42)**
Time Constraint (Days)	**-0.04 (0.01)**	**0.05 (0.02)**
Slot	**0.01 (0.004)**	**0.03 (0.01)**
Utility		
Average Daily Price (normalized)	-1	**0.11 (0.005)**
Budapest	**107.02 (54.57)**	**13.59 (1.12)**
Cancun	**146.23 (49.98)**	**21.06 (3.13)**
Manhattan	**103.16 (36.72)**	**15.14 (3.22)**
Paris	**118.58 (28.52)**	**17.60 (2.54)**
Star Rating	**37.73 (8.75)**	**12.39 (0.62)**
Consumer Rating greater than 4.5	**106.56 (23.94)**	**33.01 (13.47)**
Consumer Rating between 4 and 4.5	**78.11 (30.94)**	**13.27 (5.79)**
Distance to City Center (kilometers)	-16.92 (33.96)	**78.98 (14.27)**
Hotel Chain	**57.67 (10.04)**	**10.43 (2.58)**
Promotion Flag	**44.22 (22.14)**	10.01 (10.15)

Note: *Bold fonts indicate estimates significant at the 95% level.

Figure 1: Histogram of Consumers' Click-throughs

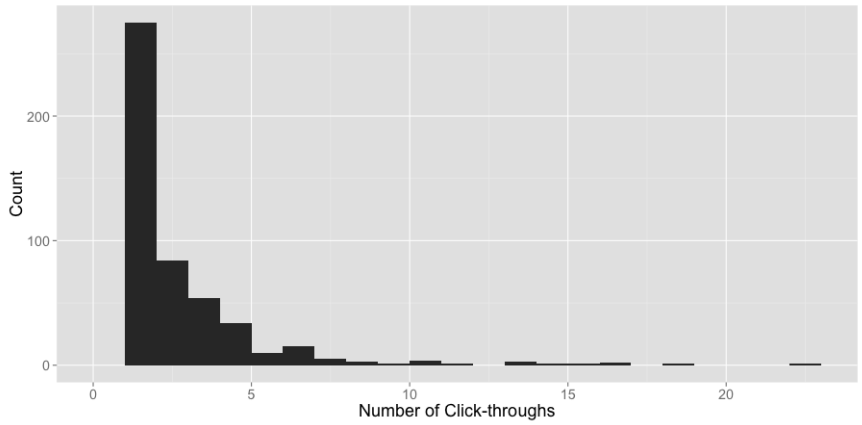


Figure 2: Histogram of Consumers' Refinement Activities (Consumers with Refinement Activity≥1)

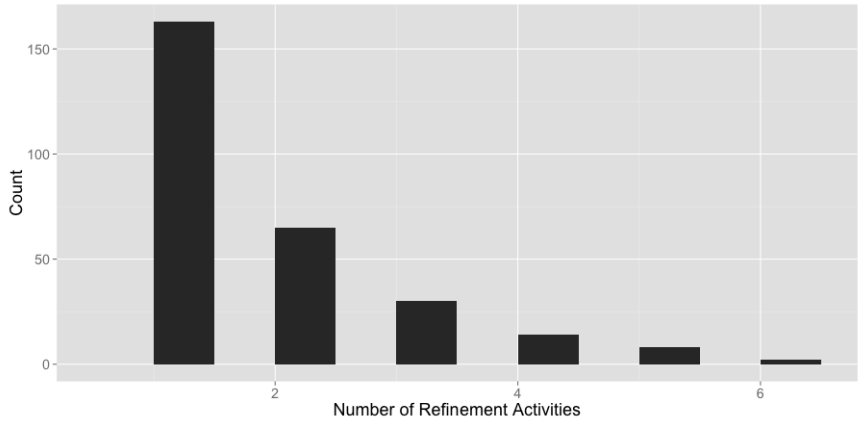
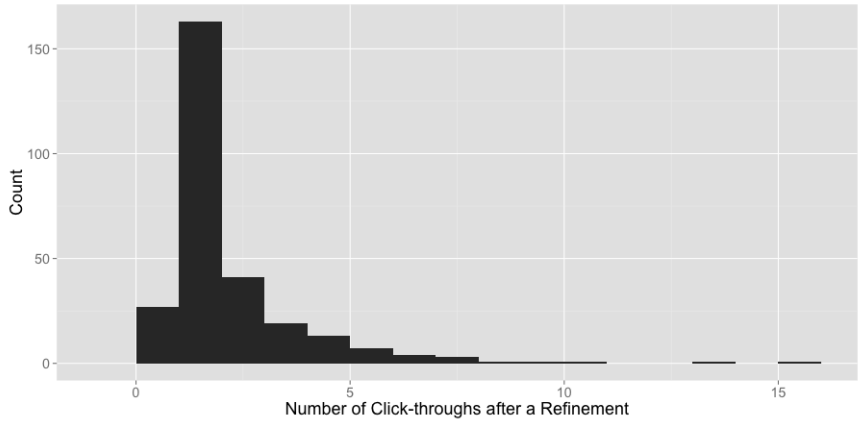


Figure 3: Histogram of Consumers' Click-throughs after Refinement Activities



References

Adam, Klaus. 2001. Learning while searching for the best alternative. *Journal of Economic Theory* **101**(1) 252 – 280.

Allenby, Greg M., Thomas S. Shively, Sha Yang, Mark J. Garratt. 2004. A choice model for packaged goods: Dealing with discrete quantities and quantity discounts. *Marketing Science* **23**(1) 95–108.

Ansari, Asim, Carl F Mela. 2003. E-customization. *Journal of Marketing Research* **40**(2) 131–145.

Arora, Neeraj, Greg M. Allenby, James L. Ginter. 1998. A hierarchical bayes model of primary and secondary demand. *Marketing Science* **17**(1) pp. 29–44.

Berry, Steven, James Levinsohn, Ariel Pakes. 1995. Automobile prices in market equilibrium. *Econometrica* **63**(4) 841–890.

Chan, Tat Y., Young-Hoon Park. 2014. The value of consumer search activities for sponsored search advertisers. *Working Paper* .

Chiang, Jeongwen. 1991. A simultaneous approach to the whether, what and how much to buy questions. *Marketing Science* **10**(4) pp. 297–315.

Chintagunta, Pradeep, Harikesh Nair. 2011. Discrete choice models of consumer demand in marketing. *Marketing Science* **30**(6) 977–996.

Chintagunta, Pradeep K. 1992. Estimating a multinomial probit model of brand choice using the method of simulated moments. *Marketing Science* **11**(4) pp. 386–407.

Chintagunta, Pradeep K. 1993. Investigating purchase incidence, brand choice and purchase quantity decisions of households. *Marketing Science* **12**(2) pp. 184–208.

Chintagunta, Pradeep K. 1998. Inertia and variety seeking in a model of brand-purchase timing. *Marketing Science* **17**(3) pp. 253–270.

De los Santos, Babur, Ali Hortacsu, Matthijs R. Wildenbeest. 2013. Search with learning. *Working Paper* .

De los Santos, Babur I., Ali Hortacsu, Matthijs R. Wildenbees. 2012. Testing models of consumer search using data on web browsing and purchasing behavior. *American Economic Review* **102**(6) 2955–2980.

Ellison, Glenn, Sara Fisher Ellison. 2009. Search, obfuscation, and price elasticities on the internet. *Econometrica* **77**(2) 427–452.

Ghose, Anindya, Panagiotis G. Ipeirotis, Beibei Li. 2012. Designing ranking systems for hotels on travel search engines by mining user-generated and crowd-sourced content. *Marketing Science* **31**(3) 493–520.

Ghose, Anindya, Panagiotis G. Ipeirotis, Beibei Li. 2014. Examining the impact of ranking on consumer behavior and search engine revenue. *Management Science* forthcoming.

Guadagni, Peter M., John D. C. Little. 1983. A logit model of brand choice calibrated on scanner data. *Marketing Science* **2**(3) pp. 203–238.

Hong, Han, Matthew Shum. 2006. Using price distributions to estimate search costs. *Rand Journal of Economics* **37**(2) 257–276.

Honka, Elisabeth. 2012. Quantifying search and switching costs in the u.s. auto insurance industry. *Rand Journal of Economics* forthcoming.

Honka, Elisabeth, Pradeep K. Chintagunta. 2013. Simultaneous or sequential? search strategies in the u.s. auto insurance industry. *Working Paper* .

Hortacsu, Ali, Chad Syverson. 2004. Product differentiation, search costs, and competition in the mutual fund industry: A case study of S&P 500 index funds. *Quarterly Journal of Economics* **119**(2) 403–456.

Kim, Jun, Paulo Albuquerque, Bart J. Bronnenberg. 2010. Online demand under limited consumer search. *Marketing Science* **29**(6) 1001–1023.

Koulayev, Sergei. 2013a. Estimating demand in online search markets, with application to hotel bookings. *Rand Journal of Economics* forthcoming.

Koulayev, Sergei. 2013b. Search with dirichlet priors: estimation and implications for consumer demand. *Journal of Business & Economic Statistics* **31** 226–239.

McDevitt, Ryan C. 2013. "A" business by any other name: Firm name choice as a signal of firm quality. *The Journal of Political Economy* forthcoming.

Mehta, Nitin, Surendra Rajiv, Kannan Srinivasan. 2003. Price uncertainty and consumer search: A structural model of consideration set formation. *Marketing Science* **22**(1) 58–84.

Moraga-Gonzalez, Jose Luis, Zsolt Sandor, Matthijs R. Wildenbeest. 2012. Consumer search and prices in the automobile market. *Working Paper* .

Nair, Harikesh, Jean-Pierre Dubé, Pradeep Chintagunta. 2005. Accounting for primary and secondary demand effects with aggregate data. *Marketing Science* **24**(3) 444–460.

Nelson, Phillip. 1970. Information and consumer behavior. *The Journal of Political Economy* **78**(2) pp. 311–329.

Petrin, Amil, Kenneth Train. 2010. A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research* **47**(1) 3–13.

Seiler, Stephan. 2013. The impact of search costs on consumer behavior: A dynamic approach. *Quantitative Marketing and Economics* **11**(2) 155–203.

Sorensen, Alan T. 2000. Equilibrium price dispersion in retail markets for prescription drugs. *Journal of Political Economy* **108**(4) pp. 833–850.

Weitzman, Martin L. 1979. Optimal search for the best alternative. *Econometrica* **47**(3) 641–654.

Yao, Song, Carl F. Mela. 2008. Online auction demand. *Marketing Science* **27**(5) 861–885.

Yao, Song, Carl F. Mela. 2011. A dynamic model of sponsored search advertising. *Marketing Science* **30**(3) 447–468.