Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

Yuxin Chen, Song Yao

To cite this article:
Yuxin Chen, Song Yao (2017) Sequential Search with Refinement: Model and Application with Click-Stream Data. Management Science 63(12):4345-4365. https://doi.org/10.1287/mnsc.2016.2557

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

MANAGEMENT SCIENCE
Vol. 63, No. 12, December 2017, pp. 4345–4365
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

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

Yuxin Chen,[a] **Song Yao**[b]

[a] New York University Shanghai, 200122 Shanghai, China; [b] Kellogg School of Management, Northwestern University, Evanston, Illinois 60208

Contact: yc18@nyu.edu (YC); s-yao@kellogg.northwestern.edu (SY)

Received: September 27, 2014
Revised: March 31, 2015; November 24, 2015; April 6, 2016
Accepted: May 6, 2016
Published Online in Articles in Advance: September 28, 2016

https://doi.org/10.1287/mnsc.2016.2557

Abstract. 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 33% more searches and enhance the utility of purchased products by 17%. Most websites by default rank search results according to their popularity, quality, 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 total search cost outweighs the enhanced utility. In contrast, if the website simply informs consumers that the default ranking already reflects product popularity, quality, or relevance, consumers search less and their surplus improves. We also find that refinement tools lead to a less concentrated market structure.

History: Accepted by J. Miguel Villas-Boas, marketing.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2016.2557.

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

1. Introduction

According to a 2011 report by McKinsey, the annual global value of search technology has surpassed $780 billion with $540 billion direct contribution to global gross domestic product since 2010 (Bughin et al. 2011). The advance of online search technology has made profound impacts on consumer behavior. In particular, search technology has helped consumers to easily form consideration sets 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 2014). 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 similarly 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, which are often used to study consumer purchase behavior (e.g., Moe 2006).

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. (2015) also use aggregate data and estimate a simultaneous search model (search with fixed

consideration set size) 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 (2014) and Honka and Chintagunta (2017) come closest to our proposed model. In Koulayev (2014), 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 (2017) identify consumers' search methods (sequential versus 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 identifying 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. Refinement tools give consumers the ability to obtain product assortments that better accommodate their heterogeneous preferences. Using the same hotel searching example, a price-sensitive consumer may use price-sorting tool to gain easier access to low price products. Another consumer looking for luxury hotels can use filters to narrow the hotels down to five-star only. In either case, the consumer only needs to focus on those options that matter the most according to her heterogeneous preference, hence finding better alternatives. More importantly, she can now place more preferable alternatives at easy-to-find positions. Due to website design, some of those alternatives' default positions are difficult to access, making them too costly to search without the refinement tools. As a result, the refinement has crucial implications on the consumer's search and purchase behavior.[1] Given the ubiquitous use of such refinement tools in online search, it is surprising that few empirical studies focus 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 33%. Furthermore, on average a consumer may achieve 17% higher utility for the product she chooses. It is crucial that websites educate consumers about their practice of ranking search results according to their popularity, quality, or relevance to consumers (e.g., Google). We find that, when consumers are uninformed about such default ranking rules, they mistakenly perceive the top ranked products on the default list have lower utility levels than they actually do. As a result, the consumers engage in disproportionately excessive searches using refinement tools, in comparison to the case when refinements are disabled. The cost accrued during the search process outweighs the utility improvement of the purchased product. Consequently, uneducated consumers' overall surplus actually drops with the presence of refinement tools. The welfare loss due to excessive searches can amount to −1.7%. 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, which ranks hotels based on consumer preferences. We show that by incorporating our model results into the ranking rule, consumer welfare can be further improved by 1.4%. This is consistent with the findings of Ghose et al. (2012 and 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, 2017), Honka (2014), Seiler (2013), and Chan and Park

(2015) propose structural models for the formation of consideration sets as the result of consumer search, and model consumer purchase conditional on the consideration sets. While Kim et al. (2010, 2017) adopt sequential search assumption, Mehta et al. (2003), Seiler (2013), and Honka (2014) consider simultaneous search assumption. Chan and Park (2015) 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 (2014) and Chan and Park (2015). A distinguishing feature of Honka (2014) and Chan and Park (2015) is their observation of consumers' consideration sets. However, neither data sets contain information of search sequences. Chan and Park (2015) 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 3. In Section 4, we discuss the identification and present some Monte Carlo simulation results. We use a click-stream data set 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 a click-through. 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 as in Weitzman (1979). Before the search starts, the consumer knows the utility level of the outside option of not buying and the expected utility of all options. At a given point during the search, suppose that the consumer has already searched some options, the consumer needs to decide (1) whether she should continue with an additional search, (2) if yes, which option 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.

Next, we consider a website as a vehicle to further fix the idea and enhance the exposition of the model. We apply the model to the click-stream data from this website, as an instantiation to demonstrate the more general applicability of the model.

2.1.1. The Website.

The website is a major U.S. online provider of travel products. We focus on consumers who are interested in booking a hotel. When visiting this website, a 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 (e.g., the previous week). 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 (frequency asked questions) 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.).[3] 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.

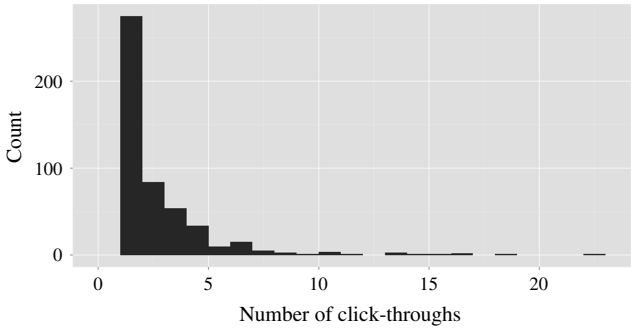
2.1.2. Consumers. The data set contains 495 individual hotel-shopping consumers' click-stream data between October 1 and October 15, 2009, for their hotel search and purchase activities. Each consumer searched hotels in one of the four cities: Budapest, Cancun, Manhattan, and Paris; and each booked one 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 data set 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 1 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.

Table 1. 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

Figure 1. Histogram of Consumers' Click-Throughs



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 2 shows the summary statistics.

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 four-star, (4) filter out hotels below three-star, (5) filter out hotels below three-star and sort by price ascendingly, (6) filter out hotels below five-star, and (7) filter out hotels below four-star and sort by price ascendingly. We group the default list and all the other less-used refinement methods as the eighth option, "(0) no refinement."

Table 2. Summary Statistics of Consumers Lapse Between the Search and the Check-In

	Mean	Std. dev.	Min.	Max.
Days between the search and the check-in (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 3. 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

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 3 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.[4] 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.

2.1.3. 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 4 reports summary statistics of hotel attributes overall and among clicked hotels.

2.2. Model

We follow Weitzman (1979) to model consumer's sequential search behavior. In Section 2.2.1, we first specify consumer utility and the effect of refinement on expected utility. We elaborate on search cost in Section 2.2.2. We then explain the optimal search strategy according to Weitzman (1979).

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

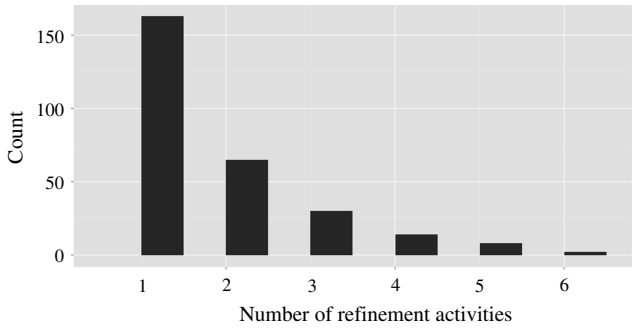
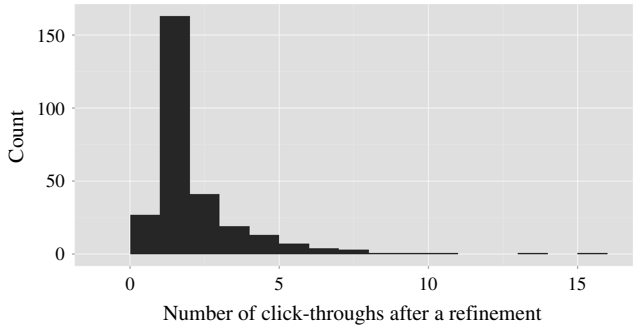


Figure 3. Histogram of Consumers' Click-Throughs After Refinement Activities



2.2.1. Utility and Refinement. Let j be the index of the sequence in which consumer i searches. Consumer i's utility of buying the product that is searched during the jth search is characterized as

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

where x_{ij} is a vector of product attributes (e.g., hotel prices and consumer ratings). x_{ij} may vary across consumers. For example, product attribute levels, such as price, may depend on when the consumer makes

Table 4. 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 ($)	252.65	96.65	233.84	76.29
Star rating	3.21	0.90	3.09	0.78
Consumer rating	3.54	1.45	3.59	1.31
Distance to city center (km)	2.15	1.74	2.01	1.62
Hotel chain	0.43	0.50	0.39	0.49
Promotion flag	0.31	0.46	0.46	0.50
Paris (1,092 hotels)				
Average daily price ($)	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

the inquiry. v_{ij} is an independent and identically distributed (i.i.d.) idiosyncratic preference shock.

Product attributes x_{ij} is drawn from some joint attributes distribution, which is assumed to be continuous and has full support. The model can easily account for the refinement as described in Section 2.1.2 (e.g., sort by price, filter by star ratings, etc.) because we may allow the attributes distribution of each search to depend on refinement, which will then affect the expected utility. To be specific, the travel website ranks the hotels list in some default order. For the jth search, the consumer needs to decide *which slot* under *which refinement tool* on the list to search. With refinement, the consumer can change the distribution of x_{ij} of the jth search in the following ways:

- *The effect of sorting on the distribution of product attributes*. If the consumer sorts the alternatives based on some attribute such as price, the sorted attribute becomes an ordered statistic. For example, if the products are sorted by price ascendingly, then products on lower slot positions *on average* have higher prices than the one on the first slot.[5] Since other attributes are correlated with price, the sorting will also have an impact on the distribution of those attributes across slots.
- *The effect of filtering on the distribution of product attributes*. Filtering on a specific attribute eliminates alternatives that do not meet the criterion. As a result, the filtering changes the attribute distribution of the listed products. 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 distribution of other attributes across slots.

To accommodate these effects of refinement, we allow the attributes distribution of the jth search to be sorting/filtering- and slot-specific. More precisely, for a given slot position $Slot_j$ and a given sorting/filtering method k ($k = 0, 1, 2, \ldots, 7$), let $P_j(x_{ij}) = P^k(x_{ij} \mid Slot_j)$ be the attributes distribution of the jth search. Essentially, the consumer's decision of *slot position and refinement for each search* affects the distribution of attributes. For example, during the jth search, suppose the consumer decides to search slot 2 using refinement method 1, then the attributes distribution of the jth search is $P_j(x_{ij}) = P^{k=1}(x_{ij} \mid Slot_j = 2)$. Such a refinement-slot-specific attributes distribution can flexibly capture the effect of refinement on search decisions and outcomes. In Online Appendix A7, we use a heuristic numerical simulation to further demonstrate the effect of the distribution specification.

For the default ranking, however, we assume that consumers treat the distribution as independent of slot position, i.e., $P^{k=0}(x_{ij} \mid Slot_j) = P^{k=0}(x_{ij})$. This is because even though the default ranking is based on past sales, the website does not inform consumers or even obfuscate the default ranking rule (see Section 2.1.1). We hence assume that when consumers view the default list, they treat the attributes distributions independent from slot positions. We test the validity of this assumption in Section 5.2.3.

In practice, we may use the empirical distribution of attributes as $P^k(x_{ij} \mid Slot_j)$. In our application, for each city we pool data together across individual consumers. For each consumer's hotel list presented by the website, we use each refinement tool to rank the hotels. Then for a slot position and a refinement method, the observed attributes across all hotels on that slot (i.e., across the lists of all consumers) can be used as the empirical distribution from which we construct $P^k(x_{ij} \mid Slot_j)$.

2.2.2. Search Cost.

The consumer incurs a search cost for each additional search. The search cost can be interpreted as resulted from 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 2014, Pinna and Seiler 2015). 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 affect the accessibility, the search cost may depend on the slot position of that search (Ansari and Mela 2003, Yao and Mela 2011, Ursu 2016).[6]

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

- The consumer knows the search cost for the jth search, c_{ij}.
- The consumer knows the distribution of product attributes $P_j(x_{ij}) = P^k(x_{ij} \mid Slot_j)$.[7]
- The consumer knows the distribution of v_{ij}.
- The consumer has a rational expectation about the expected utility of the search.[8]

2.2.3. 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 jth search, denote the cumulative distribution function of u_{ij} as $F(u_{ij})$. The distribution $F(u_{ij})$ depends on the distribution of v_{ij} and the distribution of x_{ij}, $P_j(x_{ij}) = P^k(x_{ij} \mid Slot_j)$. The expected marginal net gain from making the jth 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 incurs the search cost for the jth search.

2.2.4. 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}, and (2) making the jth 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 option with the highest utility among those already searched.

One subtlety of incorporating refinement into the model is that, under alternative refinement tools, the same hotel may appear at several different slot positions (i.e., the same hotel can appear under different search options). However, this has little impact on the consistency of our search model with the classical Weitzman's model. In our model, the same hotel under different refinement methods are essentially different search options. After one option has been searched, remaining options are still unknown to the consumer because the consumer does not know in advance which hotel will appear in the slot position of her next search. In addition, note that the distribution of attributes is search option specific. After each search, it is the search option rather than the hotel that is dropped from the remaining unsearched options. Consequently, after each search, the distribution of any unsearched option remains stable.

Practically, the reappearance of the same hotel in one's search path seldom occurs in the data used in our empirical application. This is because a hotel's slots normally change dramatically across alternative refinement methods, often many webpages apart. In our data, across different refinement tools, the average page distance of the same hotel is 3.4 pages.[9] Even for the market of Cancun, which has the least number of hotels (106), the distance is greater than one page. And as documented by multiple industrial studies, consumers mainly concentrate on the first page of search results.[10] As a result, a hotel seldom reappears at a consumer's search path.[11] Such a setting is common at online marketplaces where a consumer often faces hundreds of related products that spread across many webpages. For settings that have many reappearances of the same product, however, our model may become less applicable (e.g., when a website only has a handful of products). Such repeated searches of the same product may be related to gradual search, where a consumer learns some new information about a product during each additional search on that product (e.g., Branco et al. 2012, 2016; Ke et al. 2016).[12] A model that is constructed under a gradual search framework should be considered instead so as to better reflect such activities.

2.2.5. 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 because of the following reasons:

1. 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; and 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. In the data we observe time stamps of actions of refinement and clicks. By design, the action of a refinement proceeds the action of a click. However, the sequence of the actions may not coincide with that of the decisions. For example, as later discussed in Section 5.1.2 and shown in estimation results (Section 5.2.1), a lower slot position leads to higher search cost. For a consumer with high baseline search cost, it is possible that she decides

to search only a few top slots even before the search starts. This is because in advance she knows going down the list will be too costly for her. 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 order of these two decisions, a single-level model becomes a more natural choice.

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

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

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 is a diagonal matrix and σ_u is a column vector, both with the same dimension as α. $\xi_i\sigma_u$ 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 the diagonal elements of ξ_i follows standard normal distribution, $\mathrm{diag}(\xi_i) \sim N(0, I)$. σ_u then captures the magnitude of the heterogeneity. The random error term v_{ij} follows an i.i.d. standard normal distribution.

3.2. Heterogeneity in Search Cost

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

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

where γ are mean levels of cost coefficients. ζ_i is a diagonal matrix and σ_c is a vector. $\zeta_i\sigma_c$ are the deviations from the mean levels and capture the heterogeneity in search cost. The diagonal elements of ζ_i follows standard normal distribution such that $\mathrm{diag}(\zeta_i) \sim N(0, I)$ 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 2.2.4. 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

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

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

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

3.3.2. Search Likelihood. Denote S_{ij} as the set of options searched before the jth search. According to the optimal search strategy, when the consumer engages in the jth search, the reservation utility of that search option exceeds the realized utilities of all searched products and is the maximum among those unsearched. Furthermore, note that the reservation level of each search option depends on its attributes distribution, which is determined by the slot and refinement method. Accordingly, the likelihood function incorporates the decision of refinement. The corresponding likelihood function is

$$\begin{aligned} L_{ij}^{\text{search}} &= \Pr(R_{ij} \geq u_{ir},\ \forall\, r \in S_{ij}) \cdot \Pr\left(R_{ij} \geq \max_t R_{it},\ \forall\, t \notin S_{ij}\right) \\ &= \prod_{\forall\, r\in S_{ij}} \Pr(R_{ij} \geq u_{ir}) \cdot \Pr\left(R_{ij} \geq \max_t R_{it},\ \forall\, t \notin S_{ij}\right) \\ &= \prod_{\forall\, r\in S_{ij}} \Pr(R_{ij} \geq x'_{ir}\alpha + x'_{ir}\xi_i\sigma_u + v_{ir}) \\ &\quad \cdot \Pr\left(R_{ij} \geq \max_t R_{it},\ \forall\, t \notin S_{ij}\right) \\ &= \prod_{\forall\, r\in S_{ij}} \Pr(R_{ij} - (x'_{ir}\alpha + x'_{ir}\xi_i\sigma_u) \geq v_{ir}) \\ &\quad \cdot \Pr\left(R_{ij} \geq \max_t R_{it},\ \forall\, t \notin S_{ij}\right), \end{aligned} \tag{8}$$

$$L_i^{\text{search}} = \prod_{\forall\, j\in S_i} L_{ij}^{\text{search}}. \tag{9}$$

3.3.3. Joint Likelihood and Truncation in Preference Shocks. The joint likelihood for all consumers is therefore

$$L = \prod_i L_i^{\text{purchase}} L_i^{\text{search}}. \tag{10}$$

However, the last lines in Equations (7) and (8) do not have closed form solutions. This is because (1) both heterogeneity and revealed preference shocks of searched products (v_{ij}) are known to consumers but not observed by the researchers. They need to be integrated out when we construct the likelihood function, and (2) more importantly, the distribution of v_{ij} for those already-searched options is truncated from the perspective of researchers. To be specific,

1. Denote the final search as $\bar{j}$. For any search up to the final one, $\forall\, j < \bar{j}$, because the search continues afterward, we may conclude that realized utility u_{ij} is smaller than the reservation utilities of those searches after j. Otherwise, the search would have stopped. Furthermore, note that $R_{i\bar{j}}$ is the minimum reservation among all searched, because the search sequence is according to the ranking of reservation utilities. Hence we have

$$\begin{aligned} u_{ij} &= x'_{ij}\alpha_i + v_{ij} \le R_{i\bar{j}}, \\ v_{ij} &\le R_{i\bar{j}} - x'_{ij}\alpha - x'_{ij}\xi_i\sigma_u. \end{aligned} \tag{11}$$

2. Because the search stops after the final search $\bar{j}$, we may conclude that at least the purchased option has a utility level greater than all remaining unsearched reservation utilities. We use j^* to index the final purchase and we have

$$\begin{aligned} u_{ij^*} &= x'_{ij^*}\alpha_i + v_{ij^*} \ge \max\{R_{it},\, \forall\, t \notin S_i\}, \\ v_{ij^*} &\ge \max\{R_{it},\, \forall\, t \notin S_i\} - x'_{ij^*}\alpha - x'_{ij^*}\xi_i\sigma_u. \end{aligned} \tag{12}$$

As a result, v_{ij}'s of searched options are no longer normally distributed. From the perspective of the researchers, for searched options before the final search, the preference shocks are *right truncated*. For the purchased option, the shock is *left truncated* if it is the final search; if it is not the final search, it is *truncated on both sides*. In other words, the search sequence observed in the data determines the truncations of v_{ij}.

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 $\text{diag}(\xi_i) \sim N(0, I)$ and $\text{diag}(\zeta_i) \sim N(0, I)$, make one set of random draws 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 option (x_{ij}) in the consideration set, draw 50 sets of $\{v_{ij}\}$. Note that each v_{ij} has different truncated distribution. The truncation depends on the truncation conditions aforementioned. Repeat Steps 1 and 2 for 100 times.

3. Calculate the frequency of the condition $\{u_{ij^*} \ge u_{ij}, \forall\, j \in S_i\}$ being satisfied across the random draws of v_{ij}'s and (ξ_i, ζ_i). That is, the purchased option j^* has the highest utility level (u_{ij^*}) among all searched options. This is the simulated probability in Equation (7).

4. Calculate the frequency of the condition $\{R_{ij} \ge \max_t R_{it}, \forall\, t \notin S_{ij}\}$ being satisfied across random draws of (ξ_i, ζ_i) for each searched option j. That is, the searched option has the highest reservation among all unsearched. This is the simulated probability $\Pr(R_{ij} \ge \max_t R_{it},\, \forall\, t \notin S_{ij})$ in Equation (8).[13]

5. Repeat Steps 1–4 for all consumers to obtain the simulated likelihood for Equation (10).

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.[14] It can be shown that the following equation holds:

$$c_{ij} = \left\{(1 - \Phi(R_{ij} - \mu_{ij}))\left(\mu_{ij} - R_{ij} + \frac{\phi(R_{ij} - \mu_{ij})}{(1 - \Phi(R_{ij} - \mu_{ij}))}\right)\right\}, \tag{13}$$

where $\mu_{ij} = x'_{ij}\alpha_i$, i.e., the expected utility level with the preference shock v_{ij} integrated out. One subtle but important point is that Equation (13) relies on the shock v_{ij} being normally distributed. Unsearched v_{ij} is unknown to both the consumer and the researchers. Consequently, when the consumer evaluates the reservation utility of an unsearched option, v_{ij} does follow normal distribution. In contrast, searched v_{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 (13). 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 $\{c_{ij}, \mu_{ij}\}$.[15] We address the uncertainty of x_{ij} and hence μ_{ij} by making 30 draws repeatedly from the attributes distribution $P^k(x_{ij} \mid Slot_j)$. Based on each draw of x_{ij}, we calculate the corresponding reservation utility R_{ij}. We then average the 30 R_{ij}'s and use it as the expectation of R_{ij} if the consumer searches $Slot_j$ and uses refinement k.

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 of the interdependence between the search cost and the preference.[16] 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 (e.g., outside product has mean zero utility or price coefficient is −1),[17] 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 (v_{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 3.3.3).

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):

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

We observe purchase data, i.e., conditional on the consideration set and the truncations of preference shocks, consumers' decisions on which option 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.

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 options reveal the mean levels of preference parameters similar to classical multinomial discrete choice models.
- In search data, before the jth search, we observe (1) attributes of products up to the jth search, and (2) product attributes (exact levels or the distribution) of the jth search. Across consumers, given what have been searched in one'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 jth 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. 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 identify 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 jth 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 data sets. The first two data sets 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 products are randomly ranked. Consumers can refine the products using sorting by price or quality. 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 data sets:

- In the first data set, the search cost only has a constant term with heterogeneity. In other words, there are no exclusion restriction variables in z_{ij}. 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. We also introduce high correlation between the constant terms of the utility and the search cost, with a correlation coefficient of −0.9.
- In the second data set, 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 data sets with the same setting as the first two but with 400 consumers in each data set. Using the estimation approach proposed, Table 5 shows the results across these four simulated data sets. From Table 5, we have the following observations:

- In the first data set (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 data set (400 consumers) do not help the identification as evidenced by the insignificant estimates.
- In the second data set (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 data sets, the exclusion restrictions help to eliminate the correlation between the utility and search cost because the second data set has additional covariates that are orthogonal to the utility. More data in the fourth data set (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, without exclusion restrictions, we may not be able to separately identify the search cost from the utility.

Table 5. Monte Carlo Simulations

True values of parameters	Data set 1: Without exclusion restrictions, correlated utility, and search cost ($N = 200$)	Data set 2: With exclusion restrictions, correlated utility, and search cost ($N = 200$)	Data set 3: Without exclusion restrictions, correlated utility, and search cost ($N = 400$)	Data set 4: With exclusion restrictions, correlated utility, and search cost ($N = 400$)
Utility constant = 5	3.02 (4.33)	**4.78 (0.39)**	2.90 (2.37)	**4.84 (0.37)**
Utility constant heterogeneity = 0.5	0.71 (0.61)	**0.42 (0.09)**	0.69 (0.54)	**0.47 (0.07)**
Price= −2	**−2.92 (1.01)**	**−2.17 (0.50)**	**−2.44 (0.69)**	**−2.15 (0.29)**
Price heterogeneity = 0.5	**0.60 (0.24)**	**0.43 (0.07)**	**0.62 (0.13)**	**0.56 (0.08)**
Quality = 2	**2.51 (0.93)**	**2.24 (0.81)**	**2.47 (0.66)**	**2.23 (0.73)**
Quality heterogeneity = 0.5	**0.42 (0.12)**	**0.41 (0.13)**	**0.44 (0.13)**	**0.43 (0.08)**
Search cost constant = 2	3.80 (2.08)	**2.18 (0.35)**	2.10 (1.99)	**2.14 (0.25)**
Search cost constant heterogeneity = 0.5	0.83 (0.49)	**0.43 (0.07)**	0.71 (0.41)	**0.55 (0.05)**
Time constraint = 1	—	**0.87 (0.17)**	—	**0.91 (0.14)**
Time constraint heterogeneity = 0.3	—	**0.51 (0.12)**	—	**0.40 (0.10)**
Slot position= −1	—	**−1.20 (0.28)**	—	**−1.14 (0.20)**
Slot position heterogeneity = 0.3	—	**0.46 (0.11)**	—	**0.39 (0.10)**

Note. Bold indicates the estimates being significant at 95% level.

5. Application: Click-Stream Data of Hotel Booking

To exemplify its applicability, we apply the proposed model and estimation approach to the click-stream data set of hotel bookings provided by the travel website detailed in Section 2.1.

5.1. Utility and Search Cost Specification

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

5.1.1. Utility. The utility is specified as

$$u_{ij} = x'_{ij} \cdot (\alpha + \xi_i \sigma_u) + v_{ij}, \tag{14}$$

where x_{ij} is a vector of hotel attributes, including the hotel attributes shown in Table 4 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 v_{ij} (e.g., price is positively correlated with unobserved hotel quality contained in v_{ij}). Correspondingly, as a robustness check, we consider the control function approach (Petrin and Train 2010), treating hotel price as endogenous and using instrument variables.[18] The results are statistically indistinguishable. We do not observe systematic bias when not using the control function approach. One possibility is that there is little endogeneity after controlling the covariates included in the utility (Equation (14)). Still, in practice, the endogeneity of product attributes may be a concern and need to be evaluated carefully.

5.1.2. Heterogeneous Search Cost. To capture the heterogeneity in search cost, we specify search cost as[19]

$$\begin{aligned} c_{ij} &= c_i(\mathit{TimeConstraint}_i, \mathit{Slot}_j) \\ &= \exp(\gamma_{i0} + \gamma_{i1}\mathit{TimeConstraint}_i + \gamma_{i2}\mathit{Slot}_j) \\ &= \exp(\gamma_0 + \zeta_{i0}\sigma_{0c} + (\gamma_1 + \zeta_{i1}\sigma_{1c})\mathit{TimeConstraint}_i \\ &\qquad + (\gamma_2 + \zeta_{i2}\sigma_{2c})\mathit{Slot}_j), \end{aligned} \tag{15}$$

where $\mathit{TimeConstraint}_i$ is the number of days between consumer i's search and her check-in. Slot_j is the slot position of the jth 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. $\mathit{TimeConstraint}_i$ 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 (See Section 4.1.1).

We next discuss why $\mathit{TimeConstraint}_i$ and Slot_j are valid exclusion restriction variables, starting with time constraint. First, $\mathit{TimeConstraint}_i$ affects a consumer's search cost. Consumers only have limited amount of time to spend on searching hotels. When a consumer is closer to her travel date, time as a resource for hotel search becomes more scarce. Each unit of time spent on search is more costly. McDevitt (2014) makes a similar argument. The author shows that consumers who search for plumbers and locksmiths have higher search cost than those who search for landscapers and carpet cleaners. The reason is that consumers in the former case are often in urgent situations and hence need the services more promptly. Each unit of their time spent on searching for service providers becomes more costly. Furthermore, another similar argument is applied in Pinna and Seiler (2015), where the authors propose that people under more restrictive time-constraints have higher search cost on average. The authors use grocery stores' path-tracking data to study consumer price search behavior. For consumers who have less time for grocery shopping, each unit of time spent on price search is more costly. Also, because of their time constraints, such people tend to walk faster during grocery shopping trips. They hence use a consumer's walking speed to instrument the search cost during her grocery shopping trip. We apply a similar reasoning here and use the time till one's travel date as an exclusion restriction.

Second, $\mathit{TimeConstraint}_i$ does not enter a consumer's preference of a hotel because the hotel's utility only realizes at a later time after the consumer checks in. Accordingly, time constraint during the search does not directly affect one's utility. One concern we may have is that time constraint may affect the relative utility comparison between inside goods and outside goods. For example, if the consumer fails to book a hotel at the website because of time constraint, she may have to find one upon arrival at the destination. The inconvenience of this "outside option" makes booking a hotel at the website become relatively more appealing. In our setting, however, all consumers booked a hotel so that the relative comparison with the outside option does not affect the search behavior of the consumers in our study. Furthermore, note that the time constraint does not affect the utility comparison among the inside options. Therefore, time constraint does not affect the estimation of utility parameters in our setting.[20]

In terms of slot positions, it does not enter the utility. Again, note that the utility in the current context is derived from consuming the lodging service at a later time than the search. At the time of the consumption, slot positions at the website can no longer affect the utility. However, during search, if slot positions are correlated with unobserved hotel quality, it will affect

utility and invalidate it as an exclusion restriction variable. As a robustness check, we consider a multinomial logistic regression of consumers' final purchases on products' attributes and their slot positions. Slot position has an insignificant estimate (0.04 with a standard error of 1.26). This result implies that, conditional on other hotel attributes, slot position is conditional independent from utility, which makes it a valid exclusion restriction variable.

Furthermore, slot positions affect a consumer's search cost. Consumers often start browsing from the top of webpages as documented in the literature (e.g., Shi and Trusov 2013). The consumer needs to spend more time and efforts to reach a lower-ranked slots because of the scrolling and page-turning. It has been shown in multiple studies that high-ranked slots are more likely to be clicked, even when the models have controlled other factors such as the quality of products (e.g., Ansari and Mela 2003, Yao and Mela 2011). More recently, Ursu (2016) uses an Expedia data set to demonstrate that slot positions do affect a consumer's search cost. Expedia offers lottery-style opaque hotels to consumers so that the product can only be learnt after the purchase. The author shows that the *same* hotel is less likely to be clicked when it is demoted to a lower slot after Expedia inserts an opaque offer above it. Given that it is the same hotel, the lower number of clicks can only be attributed to search cost rather than hotel attributes.

5.2. Results and Robustness Tests

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

5.2.1. Parameter Estimates. Table 6 reports the parameter estimates. Besides city intercepts, consumer ratings have on average the highest impact on consumer utility.[21] If a hotel has a consumer rating between 4 and 4.5 stars, all other things being equal, the hotel may set its daily price \$70.01 higher than hotels with a rating lower than 4 stars. If the rating is above 4.5 stars, the premium increases to \$99.11. This is consistent with previous findings that a vendor's reputation has great impact on consumer decisions (e.g., Yoganarasimhan 2013). 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 cost of each search. 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 \$21.54 (= exp(3.07)), there are significant variations, particularly because of people's time constraints. If a consumer searches hotels 30 days in advance (the sample average), the search cost of slot 1 drops to \$4.85, in comparison to the hefty \$20.70 when the search happens on one day before the check-in (exp(3.07 − 0.05 ∗ 30 + 0.01) versus exp(3.07 − 0.05 + 0.01)). Also, if she searches slot 10 (middle of the webpage) instead of slot 1, the search cost further increases by 9%, from \$4.85 to \$5.31. Furthermore, as a heuristic way to consider the effect of refinement on search cost, the average slot of those purchased hotels under the default ranking is 19.7. With refinement tools, the average highest slot of the same set of hotels is 2.2. So the reduction in search cost is about 16%, from \$5.87 to \$4.96. Also, 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.

Table 6. Model Estimates

	Mean parameters (std. err.)	Heterogeneity (std. err.)
Search cost		
Constant	**3.07 (0.56)**	**1.10 (0.32)**
Time constraint (days)	**−0.05 (0.01)**	**0.07 (0.03)**
Slot	**0.01 (0.003)**	**0.04 (0.02)**
Utility		
Average daily price (normalized)	−1	**0.21 (0.05)**
Budapest	**114.02 (50.33)**	**15.39 (2.22)**
Cancun	**130.11 (40.00)**	**15.56 (4.83)**
Manhattan	**110.61 (36.72)**	**20.20 (5.94)**
Paris	**123.22 (38.22)**	**10.02 (4.77)**
Star rating	**40.13 (10.40)**	**18.99 (2.06)**
Consumer rating greater than 4.5	**99.11 (15.37)**	**27.02 (10.78)**
Consumer rating between 4 and 4.5	**70.01 (25.44)**	**12.37 (4.74)**
Distance to city center (kilometers)	−10.09 (26.85)	**68.98 (12.46)**
Hotel chain	**51.55 (7.04)**	**8.29 (3.32)**
Promotion flag	**48.44 (17.17)**	5.61 (7.35)

Notes. Bold indicates estimates significant at the 95% level. The standard errors are constructed using bootstrapping.

5.2.2. Model Validation. To examine the fit of the model, we consider three tests using a holdout data set. 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.[22] For each individual consumer, we can use the model estimates and the observed consumer

search/purchase data to infer her posterior parameters distribution. We then obtain 100 sets of preference and search cost parameter values, as well as 100 random utility shocks per option per consumer (v_{ij}). Conditioned on the observed hotel attributes levels, prices, and slot positions, for each set of parameter values and the random shocks, we infer which option is searched and, conditioned on the searches, which option 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 0.83 and 0.68 for search and purchase, respectively, suggesting our model captures the search behavior well.

Consumers have heterogeneous 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 another test to examine the model's ability to recover the heterogeneity. Using a similar approach, we measure the refinement methods used within each consumer's whole search process. We have 0.70 as the hit rate for the eight sorting/filtering methods (the top seven plus "no-refinement").[23]

As for in-sample fit, the corresponding hit rates of search, purchase, and refinement are 0.84, 0.71, and 0.74, respectively. Overall the model fits well.[24]

We further consider the model fit measures under an unidentified model. We remove the $TimeConstraint_i$ and $Slot_j$ from search cost function (Equation (15)). The two variables are the exclusion restrictions that separate the search cost from the utility, making the identification feasible. Table 7 shows the results of the estimation.

Table 7. Estimates of the Unidentified Model

	Mean parameters (std. err.)	Heterogeneity (std. err.)
Search cost		
Constant	4.01 (3.31)	**3.30 (1.27)**
Utility		
Average daily price (normalized)	−1	0.89 (0.60)
Budapest	95.40 (62.76)	20.61 (18.43)
Cancun	**150.32 (75.64)**	15.63 (23.31)
Manhattan	83.16 (75.88)	**10.44 (5.39)**
Pari	**100.58 (40.52)**	19.08 (20.54)
Star rating	23.31 (18.75)	**12.39 (6.62)**
Consumer rating greater than 4.5	136.50 (103.00)	13.01 (16.44)
Consumer rating between 4 and 4.5	159.11 (99.62)	**17.17 (8.90)**
Distance to city center (kilometers)	7.83 (16.10)	**78.98 (14.27)**
Hotel chain	7.01 (21.09)	**30.38 (10.87)**
Promotion flag	35.56 (29.34)	30.18 (35.49)

Note. Bold indicates estimates significant at the 95% level.

After removing the two variables, the estimates have much larger standard errors. The same set of hit rates of the model deteriorates considerably, dropping from 0.83, 0.68, and 0.70, to 0.56, 0.42, and 0.38, respectively.

5.2.3. 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.[25] This assumption implies that when the consumers view the default list, they treat the hotel attributes independent from slot positions.[26] 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:[27]

1. We reestimate 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 5.2.2. The measures change from 0.83, 0.68, and 0.70, to 0.67, 0.55, and 0.67. We take these as evidence that the original assumption (consumers do not know the rule) is more appropriate for the data.[28]

2. About 29% of consumers in the data (146 in total) can be identified as "frequent users" since they autologged 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 reestimate the model under the original assumption and the alternative one. We calculate the in-sample fit using the hit rates.[29] The hit rates under the original assumption are 0.81, 0.65, and 0.71 for these frequent users. In comparison, the measures become 0.75, 0.62, and 0.65 under the alternative assumption. This result implies that even for those who 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

$$c_{ij} = c_i(TimeConstraint_i, Slot_j) = \exp(\gamma_{i0} + \gamma_{i1} TimeConstraint_i + \gamma_{i2}\, Slot_j). \quad (16)$$

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

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

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,

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

where $1 \le Slot_j \le 25$ is the slot position on a specific webpage, and $Page_j$ is the webpage number. For example, for slot 51, $Page_j = 3$ 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 overall deteriorates as measured by out-of-sample fit (0.83, 0.68, and 0.70 versus 0.80, 0.63, 0.71; 0.83, 0.68, and 0.70 versus 0.77, 0.60, 0.61). 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 reestimate the model under two new settings using the cost function in Equation (16). 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 reestimate the model using both samples with the cost specification in Equation (16). If switching refinement methods is costly, the current cost function (Equation (16)) 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 one search after each refinement activity (136 consumers), and (2) those consumers who made at least one search after each refinement activity and more than one search after some refinement activities (146 consumers). Again, we reestimate 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.

5.3. Managerial Implications

5.3.1. Refinement and Consumer Welfare

Identified vs. Unidentified Models. To demonstrate the advantage of identifying preference and search cost parameters, we consider the following exercise. High search cost limits consumers' searches and may force them to choose options with lower utilities. As discussed earlier, search costs may be reduced when consumers are given the ability to refine search results

along the dimensions that matter the most to them. With refinement tools, high-ranking slots have more favorable attributes distribution that can accommodate a consumer's heterogeneous preference. 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 1.71 with a 95% confidence interval of $(1.43, 2.01)$.[30] With the sorting/filter options, however, each consumer on average makes 2.28 searches with a 95% confidence interval of $(2.19, 2.38)$, which represents a 33% increase in searches.[31] At the same time, the average total search cost incurred during one's search increases by only 19%, lower than the increased percentage of search numbers. More importantly, the average utility for hotels booked increases by 17% when refinement tools are available.

In Section 5.2.2, we estimate an unidentified model, where we drop the exclusion restriction variables $TimeConstraint_i$ and $Slot_j$ from search cost function. The estimates have considerably larger standard errors than the identified model. Using the estimates from this unidentified model, we consider the same exercise of simulating consumers' searches and purchases with and without the refinement tools. In this case, we find consumers on average search 1.99 times with a large 95% confidence interval of $(0.97, 3.11)$ when refinement tools are unavailable. In comparison, when refinement tools become available, the average number of searches is 2.06 with a 95% confidence interval of $(1.07, 3.20)$. In other words, there is no significant effect of refinement tools on the number of searches. Especially, between the two scenarios of with and without refinement tools, there is little difference in the total search cost incurred and the utility level of the final purchase. As a result, the website would have mistakenly undervalue the benefits of refinement tools and make suboptimal decisions on website design.

Educating Consumers About the Default Ranking Rule. As discussed above, refinement tools encourage consumers engage in more searches and result in final purchases with higher utility levels. However, it is still unclear what impact refinement tools have on the overall welfare of consumers.[32] Especially, there is a tradeoff between the amount of search and the utility of the final purchase. 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, based on the simulated searches and final purchases using the identified parameters, we further compute the net surplus of search as measured by the final purchase utility minus the total search cost incurred. Surprisingly, we find that on average the net surplus with refinement is 1.7% lower than the situation when refinements are disabled, and the 95% confidence interval of the welfare loss is $(-3.9\%, -0.3\%)$. 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. However, the consumers do not have such information about the default ranking and "mistakenly" perceive the utility levels at the high positions to be lower than they actually are. 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 using refinement tools, even though they are still making optimal decisions under the mis-information.

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 2.10, compared to 1.63 without refinement tools. The 95% confidence intervals are $(2.02, 2.19)$ and $(1.55, 1.72)$, respectively. These numbers are smaller than those made by consumers uninformed about the default ranking rule (2.28 and 1.71, respectively). We also find that net welfare surplus increases by 1.2% with the refinement, compared to the drop of 1.7% with uninformed consumers. The 95% confidence interval of the welfare improvement is $(0.7\%, 3.9\%)$. 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. Following the same fashion, we also consider a related simulation where we either keep only sorting tools or keep only filtering tools. In either case, we assume that the consumers are informed about the default ranking rule. We find that the sorting tools increase welfare surplus by 1.0% and the filtering tools increase welfare surplus by 0.8%. The surplus changes and the numbers of searches are statistically indistinguishable from each other and from the case with all refinement tools.

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 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 individual utility levels based on the model and customize the rankings of hotels individually. To investigate the effect of this alternative ranking method on consumer welfare, we consider the following policy simulation:

1. For each consumer, based on the observed search/purchase activities, infer the posterior preference parameters distribution in a Bayesian fashion.[33]

2. Conditional on observed product attribute levels and the individual's parameters posterior distribution, compute the expected utility of each hotel for the consumer.

3. Rank the hotels based on the imputed utilities in Step 2. 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 imputed utility levels.

4. Assume that consumers know this new default ranking rule. Then for each consumer, we generate 100 sets of parameters from her posterior parameters, as well as 100 random utility shocks per option.

(a) Conditioned on the observed hotel attributes levels and new positions based on the new ranking, we can simulate which hotels this consumer will search and purchase for a given set of parameters and random shocks. We can also compute the utility of the booked hotel and the net welfare of this consumer using the given set of preference parameters.

(b) To compute the *expected* utility of the booked hotel and the net welfare of this consumer, we need to integrate over the distributions of the consumer's posterior parameters and random shocks by repeating Step 4(a), using all 100 sets of parameter draws and random utility shocks, and then calculating the average of utilities.

5. We repeat Step 4 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 2.4% with a 95% confidence interval of (0.8%, 3.6%). The total net welfare of consumers increases by 2.6% with a 95% confidence interval of (0.5%, 5.0%). This simulation result implies that the new customized ranking would further enhance the overall net welfare surplus by about 1.4% for informed consumers.[34] 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 average number of searches is 2.01, lower than the average 2.10 searches under the current default ranking method with informed consumers. Although the 95% confidence intervals of the two measures overlap,[35] it is likely that the 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.

5.3.2. 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: heterogeneous consumers with refinement options.[36] Herfindahl-Hirschman Index (HHI)

is a 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.[37] We calculate search shares using the same method as in Section 5.2.2. Under the current market condition, the HHI takes the value of 0.17, 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 0.31, 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 19%, reaching 0.37. In conclusion, the refinement combined with the heterogeneity of consumers makes the market less concentrated and more competitive.[38]

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 substantial 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 data set. 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 33% more searches on average and are able to obtain 17% 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 at consumer level to determine individually customized 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

(2013) estimates a model where consumers update their Dirichlet prior beliefs on price during the search process. De Los Santos et al. (2017) 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, in real life a consumer sometimes searches the same product multiple times before purchase. One possibility is that each time the consumer gradually discovers some additional information about the product. There is a growing literature focusing on such gradual search, e.g., Branco et al. (2012, 2016), and Ke et al. (2016). While gradual search is beyond the scope of our model, it certainly reflects an important perspective of consumer search behavior and deserves more of our attention.

Third, 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. Kuksov (2004) establishes that firms may further differentiate products so as to mitigate price competition. 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. Kuksov and Villas-Boas (2010) demonstrates that the number of alternatives is an important strategic decision of firms. This is because too many or too less alternatives can both discourage consumers from search and purchase, in light of the cost of search and evaluation. Overall, we hope this paper will inspire future research on consumer online search.

Acknowledgments

This project was previously circulated under the working title "Search with Refinement." The authors thank seminar participants at City University of Hong Kong, Cornell University, Hong Kong University of Science and Technology, New York University, Ohio State University, Shanghai University of Finance and Economics, University of Chicago, University of North Carolina at Chapel Hill, Washington University in St. Louis, Xi'an Jiaotong-Liverpool University, Marketing Science Conference 2012, Third Annual Searle Conference on Internet Search and Innovation, and 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.

Endnotes

[1] In Online Appendix A6, we consider a simulation example to show how refinement may affect a consumer's decisions by lowering search cost.

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

[3] The website displays the numbers of hotels with five-star, four-star, or three-star etc. in the default search page. Therefore, the consumer knows the number of products under a filtering option before she applies that filter.

[4] We do not use the refinements without clicks in our estimation. This is a limitation of our study.

[5] When a consumer sorts the hotels based on a particular attribute, e.g., price ascendingly, the hotels may not be completely sorted according to that attribute. For example, a hotel at a higher slot may still have a higher price than a hotel at a lower slot. This is because: (1) Many websites feature certain promotional items on their lists and the refinement does not apply to those featured items. (2) Attributes (especially prices) may change during one's search duration. (3) The attribute level used by the website to sort is often different from the one after the consumer's click-through, albeit highly correlated. For example, the website uses the base price of each hotel to sort. But upon the click-through, it in fact shows the total price including different fees (amenities surcharges, taxes, city tourist fees, etc.) So there is still some uncertainty involved, even though *on average* the prices are increasing as the consumer goes down the list.

[6] In Section 5.2.3, we consider several robustness checks pertaining to the specification of search cost.

[7] We assume that consumers ex ante know the slot-refinement-specific distributions, which remain stable during the search. After one search, even for different slots under the same refinement, there is no update. Essentially, this assumption implies that there is no learning during the search. We provide robustness checks about this assumption in Online Appendix A3. The independence between search options throughout the search process is nevertheless a limitation of the model. One should check carefully with the possibility of consumers updating attributes distributions during the search. Search with learning is a crucial aspect of online shopping that deserves more attention. We call for future research on this important topic.

[8] "Rational expectation" implies a consumer's expectations about unsearched options' utility levels are unbiased. The expectations equal to the true statistical expected values and the consumer makes decisions based on these unbiased expectations.

[9] For a given hotel, the page distance is calculated based on the difference between (a) the best slot position of the hotel and (b) the average slot position of the hotel across other refinement tools. For example, if a hotel's highest position is 2 (i.e., page 1) under "sort by price ascendingly," and the average position of this hotel under the other refinement methods is 30 (i.e., page 2), then the average page distance is 1 page.

[10] For example, "The Value of Google Result Positioning," June 7, 2013, Chitika Inc., https://chitika.com/google-positioning-value.

[11] We observe that four users made 11 click-throughs on the same *hotels* (not options). We do not include such click-throughs in our

estimation (less than 1% of the click-throughs used in the estimation). We also repeat the estimation with those observations included and the results are equivalent.

[12] Furthermore, we define a search as a "click-through." If a search is more generally defined such as being exposed to the text of a link, then the gradual search framework is more applicable.

[13] Note that we do not need to calculate the component of $\prod_{\forall r \in S_{ij}} \Pr(R_{ij} - (x'_{ir}\alpha + x'_{ir}\xi_i\sigma_u) \geq v_{ir})$ in Equation (8). Condition 1 (Equation (11)) for the truncation draws in Section 3.3.3 implies the condition $(R_{ij} - (x'_{ir}\alpha + x'_{ir}\xi_i\sigma_u) \geq v_{ir})$. Consequently, $\prod_{\forall r \in S_{ij}} \Pr(R_{ij} - (x'_{ir}\alpha + x'_{ir}\xi_i\sigma_u) \geq v_{ir}) = 1$.

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

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

[16] Consumers' beliefs about the distribution of attributes may further confound the identification. Future research may relax the rational expectation assumption we adopt and model consumer learning of the distribution of attributes during the search.

[17] Without the normalization, the model may be unidentified. To see this, note that the optimal search strategy is governed by reservation utility R_{ij} of each option, which is solved from the implicit function in Equation (4), $\int_{R_{ij}}^{\infty}(u_{ij} - R_{ij})\,dF(u_{ij}) = c_{ij}$. It is possible to have multiple sets of parameters to satisfy the equation, especially when we have constant intercept terms in the utility and search cost specification.

[18] We first use instruments to predict prices. In the second step, we plug in the residual term as an additional control in the nonlinear estimation, as discussed in Petrin and Train (2010). We choose the instruments so that they affect hotel pricing decisions but are independent of the unobserved hotel quality (Berry et al. 1995, Hortacsu and Syverson 2004). In particular, we use average prices of the same market, average prices of hotels in the same market with the same star rating, average prices of hotels in the same market with the same consumer rating (all excluding the focal hotel), and own hotel nonprice attributes. Alternatively, one can include product-specific intercepts in utility specification. We do not have hotel intercepts because of data limit. In practice, however, one should try to control product-specific intercepts as data permit so as to alleviate endogeneity concern.

[19] We also consider two alternative search cost specifications in Section 5.2.3 and confirm the current specification is appropriate.

[20] In a more general case where there is an outside option, the time constraint only affects the comparison between inside and outside goods. The relative utility comparison among inside options are still independent from the time constraint. Nevertheless, when outside options are included in the model, some additional exclusion restrictions may deserve further consideration.

[21] Ideally, we may estimate each of the four markets separately because of potential unobserved market-specific heterogeneity. However, given the data size, we decide to pool all observations together to gain statistical power. We include city-intercepts to control market fixed effects.

[22] We detail the implementation in Online Appendix A5.

[23] The hit rate of refinement is calculated in the following way. For example, suppose for a given consumer, we observe that she makes three searches. The three searches use refinement method 1, method 2, and method 3, sequentially. The model predicts that she uses method 1, method 2, and method 4, sequentially. Then the hit rate of this consumer is $2/3 = 0.67$. That is, the model predicts correctly two out of three refinement methods in correct sequence used by the consumer. We then aggregate the individual hit rate across all consumers and obtain the hit rate of 0.70.

[24] As a benchmark for model fit, we estimate logit models for search, refinement, and purchase, respectively. For search, the dependent variable is which unsearched hotel will be clicked. The variables controlled are hotel attributes and default ranking positions. For refinement, the dependent variable is which refinement methods is chosen for a given click-through, depending on the averages and standard deviations of the attributes of already searched and unsearched. For purchase, the dependent variable is which hotel is booked among the searched, depending on each searched hotel's attributes. We find that the out-of-sample hit rates for search, purchase, and refinement are 0.52, 0.35, and 0.43, respectively.

[25] The website later changed the label of the default ranking from "[Website] Picks" into "Most Popular." This may be a piece of anecdotal evidence that the website wanted to make it more clear to consumer what the default ranking meant.

[26] Another assumption we have pertaining to consumer information structure is that after refinement a consumer knows slot- and refinement-specific attributes distribution $P_j(x_{ij}) = P^k(x_{ij} \mid Slot_j)$. It is possible, however, that the consumer is agnostic about the distribution after refinement. To test the validity of this assumption, we consider two additional robustness tests in Online Appendix A1.

[27] Note that these robustness checks are *necessary conditions* for the validity of the information structure assumption.

[28] We also consider a test where we randomly select 50% of consumers and assume that they know the default ranking rule. The other 50% remain oblivious to the ranking rule. We then calibrate the out-of-sample hit rates, which are 0.69, 0.58, 0.69. They are slightly better than the case of "knowing the rule" but still worse than the case of "not knowing the rule." Empirically, we cannot completely rule out the possibility that some consumers do understand the rule. The true state is probably somewhere between the two extremes. But it is very crucial for firms to enhance consumers' knowledge about their websites. We further explore this managerial issue in Section 5.3.

[29] 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).

[30] In Online Appendix A5, we detail how the simulations are implemented.

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

[32] While the empirical tests in Section 5.2.3 seem to support the assumption that the consumers do not understand the default ranking rule, the true knowledge state of consumers is potentially between full knowledge and complete obliviousness. In practice, it is very crucial for firms to enhance consumers' knowledge about the website design.

[33] See Online Appendix A5 for the details of implementation. Note that in practice, instead of the observed purchase, we can use past purchases instead.

[34] The percentage 1.4% is computed as $(2.6\% - 1.2\%)$. The benchmark 1.2% is the welfare improvement when consumers are informed about the current default ranking rule (the counterfactual considered above).

[35] The 95% confidence intervals are $(1.90, 2.13)$ and $(2.00, 2.19)$, respectively.

[36] 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.

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

[38] Note that this simulation is implemented in the spirit of a *comparative static*. While holding other things equal, we evaluate the (marginal) effect of search tools on market structure on the website. Readers should be careful of generalizing the results into the real world scenarios, where "other things equal" are unlikely to hold.

References

Adam K (2001) Learning while searching for the best alternative. *J. Econom. Theory* 101(1):252–280.

Ansari A, Mela CF (2003) E-customization. *J. Marketing Res.* 40(2): 131–145.

Berry S, Levinsohn J, Pakes A (1995) Automobile prices in market equilibrium. *Econometrica* 63(4):841–890.

Branco F, Sun M, Villas-Boas JM (2012) Optimal search for product information. *Management Sci.* 58(11):2037–2056.

Branco F, Sun M, Villas-Boas JM (2016) Too much information? Information provision and search costs. *Marketing Sci.* 35(4):605–618.

Bughin J, Corb L, Manyika J, Nottebohm O, Chui M, de Muller Barbat B, Said R (2011) The impact of Internet technologies: Search. Report, McKinsey & Company, New York.

Chan TY, Park Y-H (2015) Consumer search activities and the value of ad positions in sponsored search advertising. *Marketing Sci.* 34(4):606–623.

De Los Santos BI, Hortacsu A, Wildenbees MR (2012) Testing models of consumer search using data on web browsing and purchasing behavior. *Amer. Econom. Rev.* 102(6):2955–2980.

De Los Santos B, Hortacsu A, Wildenbeest MR (2017) Search with learning for differentiated products: Evidence from e-commerce. *J. Bus. Econom. Statist.* 35(4):626–641.

Ellison G, Ellison SF (2009) Search, obfuscation, and price elasticities on the Internet. *Econometrica* 77(2):427–452.

Ghose A, Ipeirotis PG, Li B (2012) Designing ranking systems for hotels on travel search engines by mining user-generated and crowd-sourced content. *Marketing Sci.* 31(3):493–520.

Ghose A, Ipeirotis PG, Li B (2014) Examining the impact of ranking on consumer behavior and search engine revenue. *Management Sci.* 60(7):1632–1654.

Hong H, Shum M (2006) Using price distributions to estimate search costs. *RAND J. Econom.* 37(2):257–276.

Honka E (2014) Quantifying search and switching costs in the U.S. auto insurance industry. *RAND J. Econom.* 45(4):847–884.

Honka E, Chintagunta P (2017) Simultaneous or sequential? Search strategies in the U.S. auto insurance industry. *Marketing Sci.* 36(1):21–42.

Hortacsu A, Syverson C (2004) Product differentiation, search costs, and competition in the mutual fund industry: A case study of S&P 500 index funds. *Quart. J. Econom.* 119(2):403–456.

Ke TT, Shen Z-JM, Villas-Boas JM (2016) Search for information on multiple products. *Management Sci.* 62(12):3576–3603.

Kim JB, Albuquerque P, Bronnenberg BJ (2010) Online demand under limited consumer search. *Marketing Sci.* 29(6):1001–1023.

Kim JB, Albuquerque P, Bronnenberg BJ (2017) The probit choice model under sequential search with an application to online retailing. *Management Sci.* 63(11):3911–3929.

Koulayev S (2013) Search with dirichlet priors: Estimation and implications for consumer demand. *J. Bus. Econom. Statist.* 31(2): 226–239.

Koulayev S (2014) Search for differentiated products: Identification and estimation. *RAND J. Econom.* 45(3):553–575.

Kuksov D (2004) Buyer search costs and endogenous product design. *Marketing Sci.* 23(4):490–499.

Kuksov D, Villas-Boas JM (2010) When more alternatives lead to less choice. *Marketing Sci.* 29(3):507–524.

McDevitt RC (2014) "A" business by any other name: Firm name choice as a signal of firm quality. *J. Political Econom.* 122(4): 909–994.

Mehta N, Rajiv S, Srinivasan K (2003) Price uncertainty and consumer search: A structural model of consideration set formation. *Marketing Sci.* 22(1):58–84.

Moe WW (2006) An empirical two-stage choice model with varying decision rules applied to internet clickstream data. *J. Marketing Res.* 43(4):680–692.

Moraga-Gonzalez JL, Sandor Z, Wildenbeest MR (2015) Consumer search and prices in the automobile market. Working paper, University of Amsterdam, Amsterdam.

Nelson P (1970) Information and consumer behavior. *J. Political Econom.* 78(2):311–329.

Petrin A, Train K (2010) A control function approach to endogeneity in consumer choice models. *J. Marketing Res.* 47(1):3–13.

Pinna F, Seiler S (2015) Consumer search: Evidence from path-tracking data. Working paper, London School of Economics, London.

Seiler S (2013) The impact of search costs on consumer behavior: A dynamic approach. *Quant. Marketing Econom.* 11(2):155–203.

Shi S, Trusov M (2013) The path to click: Are you on it? Working paper, Santa Clara University, Santa Clara, CA.

Sorensen AT (2000) Equilibrium price dispersion in retail markets for prescription drugs. *J. Political Econom.* 108(4):833–850.

Ursu RM (2016) The power of rankings: Quantifying the effects of rankings on online consumer search and choice. Working paper, New York University, New York.

Weitzman ML (1979) Optimal search for the best alternative. *Econometrica* 47(3):641–654.

Yao S, Mela CF (2008) Online auction demand. *Marketing Sci.* 27(5):861–885.

Yao S, Mela CF (2011) A dynamic model of sponsored search advertising. *Marketing Sci.* 30(3):447–468.

Yoganarasimhan H (2013) The value of reputation in an online freelance marketplace. *Marketing Sci.* 32(6):860–891.