RAND Journal of Economics
Vol. 45, No. 3, Fall 2014
pp. 553–575

Search for differentiated products: identification and estimation

Sergei Koulayev*

When consumers search for differentiated products, a given search decision can be explained either by low search cost or by low tastes for the set of products already found. We propose an identification strategy that allows to estimate the search cost distribution in the presence of unobserved tastes. The required data takes the form of conditional search decisions: observations of search actions combined with previously observed product displays. We develop an application using clickstream data from a hotel search platform. Estimates of price elasticity of demand in the search model differ from those in the static model, reflecting the bias due to endogeneity of search-generated choice sets.

1. Introduction

■ In markets with multiple sellers and frequently changing prices, consumers often have to engage in costly search in order to collect information necessary for making a purchase. The search results in a collection of products from which the purchase decision is made. These search-generated choice sets possess two distinct properties. First, because search is costly, they are usually limited compared to the full set of available products: according to comScore data,[1] only a third of all consumers visit more than one store while shopping online. Second, they are correlated with consumer preferences, as the decision to stop searching is in part dictated by the expected benefit of search. These properties violate key assumptions behind the commonly used approach to demand estimation (e.g., Berry, 1994; Berry, Levinsohn, and Pakes, 1995), and new techniques need to be developed in order to infer consumer preferences in search markets.[2]

* Consumer Financial Protection Bureau; sergei.koulayev@gmail.com.

Disclaimer: The views expressed in this article are author's only and do not represent the views of CFPB.

This article has grown out of my PhD dissertation at Columbia University. I am grateful to my advisors, Michael Riordan, Bernard Salanié, and Katherine Ho, for their advice and encouragement. Special thanks to Marc Rysman, Jose Moraga-Gonzalez, Paulo Albuquerque, and Arthur Lewbel for their helpful suggestions. The web Appendix for this article contains necessary derivations and can be found at www.sergeikoulayev.com/rand5appendix.pdf.

[1] As reported by de los Santos (2008), the number is 27% in 2002 and 33% in 2004. In our data, too, only a third of searchers look at more than one page of hotel options resulting from the search request. See also Johnson et al. (2004) for additional evidence on search intensity on the web.

[2] Beyond demand estimation, magnitudes of search costs are a necessary input into optimal design of search platforms, where one needs to predict how consumers will react to new search tools, changes in recommendation rankings, etc.

Central to the inference of search-generated demand is the issue of identification of search costs. As pointed out by Sorensen (2000), explaining search decisions made by consumers with heterogeneous preferences contains an identification problem. A person may stop searching either because she has a high valuation for the products already found, or because she has a high search cost. One way to solve this problem is to use exogenous shifters of search costs: for instance, Moraga-Gonzalez, Sandor, and Wildenbeest (2010) used distances to car dealerships to estimate costs of search for a new car.[3] In online markets, such data is rarely available, so alternative approaches are needed. The key advantage of web-based search is the availability of detailed logs of browsing and clicking activity, and several recent articles have used it. For example, Kim, Albuquerque, and Bronnenberg (2010) exploit the unique properties of the recommendation system on Amazon.com to identify search costs. In the context of online hotel bookings, Ghose, Ipeirotis, and Li (2012) estimate a search model where consumers are learning about the quality of a hotel brand.[4] Panel logs from comScore data are employed in de los Santos, Hortacsu, and Wildenbeest (2012) for nonparametric estimates of search costs.

Although the number of applications using online search data is growing, the issue of identification remains open. The existing results are limited to homogeneous (Hong and Shum, 2006) or vertically differentiated (Hortacsu and Syverson, 2004) products. In this article, we obtain a nonparametric identification of search cost distribution for the case of preferences with common mean utility and i.i.d logit taste shocks, a first step toward identification of search costs in the presence of taste heterogeneity. To achieve identification, the search data should be presented as a sequence of conditional search decisions. A conditional search decision includes a search action in conjunction with the set of products observed prior to that action. This set constitutes a fall-back option, or status quo, when a consumer is deciding whether to search. Changes in product displays provide an exogenous source of variation in status quo, which leads to variation in search intensity that is independent of search costs.

As an application, we estimate a dynamic model of search, using a data set of observation histories, search actions, and clicks for hotels from a major search aggregator. To reduce the burden of numerical integration, we derive analytical expressions for the likelihood of joint searching and clicking decisions. These expressions are general enough to be applicable for any sequential search for differentiated products, where data on conditional search decisions is available.

We find that search costs in this environment are large: median cost is around \$10 per page of results, and can be as high as \$30 for a subset of population. We interpret these costs as cognitive costs of comparing prices and other characteristics of newly found hotels. Another way to understand these magnitudes is to observe that many consumers choose to forego potential benefits of search, which are substantial for highly differentiated products such as hotels. Further, we find evidence of multimodality of search costs among population, with modes in \$10, \$24, and \$30 of search costs. This result calls for more flexible forms of search cost distribution than the typically used log-normal distribution.

An important phenomenon in the data is consumers returning to previously found search results. Within a rational search model, such behavior requires that reservation utilities are decreasing with each search attempt. This may occur either if consumers become more pessimistic,

[3] Also, Brynjolfsson, Dick, and Smith (2010) offer a way to approximately estimate search costs by compensating variation between a smaller (no search) and a larger (search) choice sets. However, their approach requires existing demand estimates, so that valuations can be computed.

[4] Currently, there are two separate notions of search in the literature. One is a search that leads to discovery of existing varieties (or prices), as in this article, as well as Sorensen (2000), Hortacsu and Syverson (2004), Hong and Shum (2006), and de los Santos (2008). Another is search as learning about quality of known varieties (brands), also called "consideration set formation." Ghose, Ipeirotis, and Li (2012), as well as Kim, Albuquerque, and Bronnenberg (2010) are the latest examples of this approach (earlier articles include Roberts and Lattin, 1991; Mehta, Rajiv, and Srinivasan, 2003). Both views on search are complementary as they address two different types of consumer uncertainty.

or if search costs are increasing. Allowing for the latter scenario, we find median search costs rising from $4 per first search to $16 per fifth.

Going back to the original purpose of this study, we use estimates from search model to evaluate biases due to endogeneity of search-generated choice sets. Comparing price elasticities implied by the search model to those obtained under a static discrete-choice model, we find meaningful differences (as much as 30% for some specifications), although the sign of the bias cannot be predicted. This implies that search decisions are informative about consumer preferences and need to be accounted for in the inference of consumer demand.[5]

The rest of the paper is organized as follows. Section 2 outlines the main properties of the data. Section 3 presents the model of joint clicking and searching decisions, as well as relevant data summaries. Section 4 discusses the identification of the main parameters of the model. Section 5 presents external (model-free) evidence on possible magnitudes of search costs in this environment. Section 6 presents estimation results and Section 7 concludes.

2. Data

■ A consumer is searching for a hotel in Chicago on a popular search aggregator. To begin search, she submits a search request, which includes the city (Chicago), dates of stay, number of guests, and number of rooms. An average request results in more than 140 available hotels, leading to a nontrivial search problem. When results are loaded in the browser, the visitor observes the first page of results, which contains 15 hotel options sorted by a default ranking criterion that is a part of the website's recommendation system. To explore more options, users can flip through pages of recommended hotels or employ various sorting and filtering tools.

As soon as the user finds a preferred hotel, she can click on it, which redirects the user to another website where a booking can be made. If several clicks are made (about 40% of users do so), we take the last click for the analysis. Because the click takes the user away from the website, the last click also finalizes the search session.

In total, the data set contains 23,959 unique search histories, by consumers who visited the website during May 2007. For every search history, we observe (i) parameters of the initial request (including the date of search); (ii) the sequence of search actions; (iii) displays of hotels and their prices, observed after every search action; and (iv) identities and prices of clicked hotels. This data offers a very detailed picture of the user search experience, showing how the user's information set was evolving during search, and which among discovered hotels were preferred.

For all its benefits, these data have two shortcomings. First, we observe only clicks, but not final bookings. This limitation is common to online data, as discussed by Brynjolfsson, Dick, and Smith (2010). In another study, Brynjolfsson and Smith (2001) compare clicks and actual purchases of books and conclude that a click is a valuable indicator of preferences, albeit a noisy one. Our results also lend support to this conclusion, as we find economically plausible effects of a hotel's characteristics on the click rate. Therefore, we view click as a revealed preference action, where utility of the clicked hotel is highest among hotels in the choice set. Second, because users are anonymous, we cannot connect searches made by the same person, more than 24 hours apart. In the estimation, we will take each search history as being made by a separate individual. To the extent that the possibility of future search can serve as a substitute for the current search effort, our estimates of search costs will be biased upward. This is probably less of an issue with the hotel market, where prices change rapidly, so that current results may not be available at a later date. In other words, we focus on search decisions made within a relatively short time frame of a single search session, when all previously observed results are available.

[5] Brynjolfsson, Dick, and Smith (2010) also evaluate the endogeneity bias in the price sorted environment: they estimate a static discrete-choice models on two subsamples, one of users who did not search and clicked on the first page, and another of those who turned the page to look at more expensive books. They find significantly higher price elasticity in the first group than in the second, indicating that the static model overestimates price elasticity.

□ **Search strategies.** Explaining search activity in this search environment within a rational model is a challenging task. The number of potential search strategies is very large, due to availability of sorting and filtering tools. One approach is to focus on a subset of consumers who employed a narrowly defined search strategy. This is a feasible but not preferred approach, as it does not identify the cost of the first search attempt and may give biased results. Another solution is to model every search action, but it is not feasible.

We attempt to strike a balance between these extremes. We choose a handful of most popular search strategies and model decisions of consumers who used these strategies in a fully structural way. For consumers who used other strategies take only the first search action and ignore the rest. At least, this approach gives an unbiased view on the baseline search cost associated with the first search.

Table 2 ranks search strategies by popularity. The first ”strategy” is not to search at all: after observing the first page, either click on one of those hotels, or leave the website without clicking. Because the first page of results appears immediately after the search request, it is seen by all visitors. This passive option is by far the most popular one, preferred by 35% of visitors. The second most popular strategy is to flip through pages of recommended hotels (13% of users). With each search effort, a new set of 15 hotels is revealed. The third strategy is to sort hotels by increasing price (5% of users). Following that action, the user observes the 15 cheapest hotels and may continue searching by exploring more and more expensive hotels. The fourth strategy is to sort hotels by increasing distance to the city center (1.3% of users). Combined, these four strategies account for 54% of the sample, or 12,930 unique search histories. For consumers who used more rare search strategies, we record only the first search action (as detailed on lines 5–19 in Table 2). Together, lines 1 to 19 add up to the estimation sample of 19,291 observations, or 80% of the original sample. Other consumers have made very rare or specific actions, such as searching by hotel name.

□ **Searching and clicking.** Figure 1 presents search intensities in the estimation sample. The darker areas correspond to the structural sample, and the lighter area corresponds to the first search action made by users who chose other search strategies (lines 5–19 in Table 2). We can see that a substantial part of search activity happens outside the structural sample, so it is important to account for this activity in the inference of search costs. Further evidence is presented in Table 3. Conditional on the length of search, we report the number of observations, the click rate, and the contribution to the total number of clicks. Consumers who search actively also click more: the click rate among passive users is 29.6%, and among those who made a search effort, it is 34.3%. Accordingly, searchers bring disproportionate shares of total clicks: being 35.8% of the sample, they bring 38.2% of clicks. This fact is consistent with the discrete–choice model, which stipulates that consumers who observed more options (as a result of search) are more likely to click.

□ **Chicago hotels.** In total, 148 various Chicago hotels were displayed to users who searched during May 2007. Because we do not observe the total availability for each request, we assume that all 148 hotels were available.[6] They are located in the city of Chicago itself, in satellite towns (Evanston, Skokie, etc.), or in the proximity of airports (O'Hare, Midway–see the web Appendix[7] for a map of hotel locations). For each hotel, search results display its name, chain affiliation, price, star rating, neighborhood, and distance to the city center. Although additional information can be collected, in the estimation we use only characteristics that were displayed to the user, as they were likely to have the most immediate effect on the click rate.

[6] This assumption is relevant for the specification of a consumer's beliefs, and only in this way it affects our results. We checked the availability by entering search requests for various dates, and found that it did not vary much.

[7] At www.sergeikoulayev.com/rand5appendix.pdf.

□ **Price variation.** The relevant product definition in this market is a one night stay at the specified date in the future. If several room types are available, the website shows the cheapest among them. Fortunately for us, the hotel market exhibits significant price fluctuations: the average price is $230 with standard deviation of $127. There are three sources of price variation: first, differences in quality characteristics among hotels; second, changes in the attractiveness of a given hotel over time; third, changing inventory of rooms available at a given hotel for a future date. The second and the third factors imply that if two users search on different days, or for different arrival dates, they may see different prices for the same hotel.

We also find that about 14% of within-hotel price variation cannot be explained by the observable factors: searchers with the same combinations of date of search and date of arrival are shown different prices for the same hotel (see the web Appendix for more details). Our hypothesis is that hotels or online travel agents (OTAs) are engaged in a sort of ”experimental” pricing, where they randomly change prices in order to capture some of the high-value consumers.

□ **Request types.** Parameters of search request include date of search, dates of stay, number of people, and number of rooms. From the dates of search and stay, we derive advance purchase, length of stay, and whether Saturday night is included. Table 5 presents summary statistics. The median advance search in our sample is 21 days, and 60% of users plan to stay over Saturday night. They often travel in groups of two or more people (the median is two persons). In our analysis, we combine parameters of request into a number of "types," which may reflect underlying characteristics of the consumer. For example, people who search further in advance are expected to produce less clicks, as they have more chances to search later.

□ **First page variation.** The first page of results, observed by all users, provides a fall-back option, or status quo, against the option of searching further. As we will explain in more detail below, the identification strategy requires variation in contents of that page across users. Table 6 presents summary statistics of price, star rating, distance to center, and distance to O'Hare, by position on the first page. The standard deviations of these attributes reveal substantial heterogeneity among hotels that appear on any position. With respect to star rating, the variation is smaller: typically, these are three- or four-star hotels, with occasional two-star or five-star hotel. See the web Appendix for additional information.

3. Model

■ In many ways, our search model is motivated by the specifics of the search environment. This approach helps avoid making ad hoc assumptions regarding the search process that other studies have relied on.

The search process unfolds in two stages, as illustrated by Figure 2. Initially, the user submits a search request and observes the first page of results, whose contents are chosen by the website's recommendation system. The search results page displays 15 hotels: name, price, and a limited set of attributes. In addition, every hotel is associated with a consumer-specific match value that is unobserved by the econometrician. We assume that consumers learn prices, attributes, and match values of hotels on the first page at no cost. Conditional on this information, consumer can either click on one of the first page hotels or continue searching. A click terminates the search process, and so does the choice of the outside option, when the consumer leaves the website without clicking. If she decides to continue searching, there are several search strategies available, as detailed in Table 2.

After a search strategy is chosen, the user makes a sequence of search attempts. After each attempt, she has two options: continue searching or click on one of the previously observed hotels. If she decides to search, she pays the search cost and learns prices, attributes, and match values of 15 hotels on the next page. If the user reaches the terminal page, she cannot search further but must click on one of the previously observed hotels, or leave without clicking (we assume the user can search at most six pages).

Two features make search on this website different from the standard sequential search model. First, the search is nonstationary. Because hotels are always ordered in some way, beliefs about the distribution of offers change with each search attempt, in a nonrecursive fashion. Second, the search horizon is finite. Both features imply that search decisions are forward looking: they incorporate both immediate benefits of search and the option of searching further. In contrast, the standard sequential search model is a static one, as search decisions are dictated only by the immediate benefit of search.

□ **Optimal search: reservation utility characterization.** We begin by considering the optimal search length given the choice of search strategy. Let u_{1i} be the highest utility on the first page, or simply the utility of that page, for consumer i. There is also an option of leaving the website without clicking. Its value is u_{i0} and it is available at any time during search. Accordingly, the status quo after the first page is the best between these two options, denoted as $u^*_{1i} = \max\{u_{0i}, u_{1i}\}$. If the user continues to search, she observes the second page of results, with maximal utility of u_{2i}. Following that, her status quo becomes $u^*_{2i} = \max\{u_{0i}, u_{1i}, u_{2i}\}$. More generally, let index t indicate the number of pages already observed. Then, the status quo in period $t = 1..T$ is given by $u^*_{ti} = \max\{u_{0i}, u_{1i}, .., u_{ti}\}$. In a recursive way, the evolution of the status quo can be expressed as:

$$u^*_{ti} = \max\{u_{ti}, u^*_{(t-1)i}\}. \tag{1}$$

Search costs may change by period. The first period search cost is denoted by c_{1i}, it is the cost of exploring the second page of results (the first page is for free). We also call it baseline search cost. Index i indicates that every consumer is endowed with idiosyncratic value baseline search cost, as a draw from some distribution. Future search costs are given by:

$$c_{ti} = c_{1i} + \Delta_t,\ t = 2..T-1. \tag{2}$$

As search progresses, so do beliefs regarding contents of the subsequent page. For example, with price sorting strategy, the user expects more expensive hotels with every next page. Let $\tilde{u}_{t+1}$ represent the best utility of the next page, a random quantity. Then, $G_t(\tilde{u}_{t+1}|\theta_i)$ is the distribution of search results, conditional θ_i - vector of consumer characteristics. Because the evolution of beliefs cannot be expressed in a recursive fashion, time becomes a state variable in this model.

To summarize, there are three state variables: status quo, u^*_{ti}, search cost, c_{ti}, and the time index t. The value function in period t is:

$$V_t(u^*_{ti}, c_{ti}|\theta_i) = \max\{u^*_{ti}, E_t\left(V_t(\tilde{u}^*_{t+1}, c_{(t+1)i}|\theta_i)|u^*_{ti}\right) - c_{ti}\}. \tag{3}$$

Together, equations (1), (2), and (3) characterize the dynamic problem faced by the consumer. The expectation operator in (3) is obtained from the current distribution of beliefs as:

$$E_t\left(V_t(\tilde{u}^*_{t+1}, c_{(t+1)i}|\theta_i)|u^*_{ti}\right) = \int V_{t+1}\left(\max\{\tilde{u}_{t+1}, u^*_{ti}\}, c_{(t+1)i}|\theta_i\right) dG_t(\tilde{u}_{t+1}|\theta_i).$$

As consumers do not face any uncertainty about search cost innovations, the period search cost $c_{(t+1)i}$ can be included as a part of the vector θ_i. As a result, the expectation operator reduces to:

$$Q_t(u^*_{ti}|\theta_i) = E_t\left(V_t(\tilde{u}^*_{t+1}|\theta_i)|u^*_{ti}\right) \tag{4}$$

$$= E_t \max\left(\max\{\tilde{u}_{t+1}, u^*_{ti}\}, Q_{t+1}\left(\max\{\tilde{u}_{t+1}, u^*_{ti}\}|\theta_i\right) - c_{t+1i}\right). \tag{5}$$

where in the second step, we used the law of motion for status quo.

We solve for the optimal policy function using backward induction. This method requires the terminal period, which we choose as $T = 6$, a generous limit given our data. Let $t = T - 1$, so that only one search attempt remains. The continuation value of search (gross of search costs) is:

$$Q_{T-1}(u^*_{T-1i}|\theta_i) = E_t \left(\max\{\tilde{u}_T, u^*_{T-1i}\}|\theta_i\right). \tag{6}$$

If the distribution of search results, $\tilde{u}_T$, has full support,[8] then it is straightforward to verify that there exists a critical level of status quo that makes the consumer indifferent between searching and not searching: $u^*_{T-1i} = Q_t(u^*_{T-1i}|\theta_i) - c_{(T-1)i}$. The value of u^*_{T-1i} that solves this equation is called reservation utility and denoted by $r_{(T-1)i}$. Alternatively, the reservation utility can be found from the equation:

$$Q_{T-1}\left(r_{(T-1)i}|\theta_i\right) - r_{(T-1)i} = c_{(T-1)i}. \tag{7}$$

which simply states that the expected benefit of search is equal to the search cost. Accordingly, the consumer i will search in period $T-1$ if and only if $u^*_{T-1i} < r_{(T-1)i}$. More generally, for any period $t \leq T-1$, the reservation value is given as a solution to the equation:

$$Q_t\left(r_{ti}|\theta_i\right) - r_{ti} = c_{ti}. \tag{8}$$

After observing $t < T$ pages of results, a consumer will search if and only if:

$$u^*_{ti} < r_t(c_{ti}). \tag{9}$$

The vector of reservation utilities, $r_{1i}, .., r_{T-1i}$ completely determines the search behavior: consumer i will optimally search t_i pages of results if and only if all inequalities hold:

$$u^*_{1i} < r_{1i}, .., u^*_{(t_i-1)i} < r_{(t_i-1)i}, u^*_{t_i i} > r_{t_i i}. \tag{10}$$

The functions Q_t, $t = 1..T-1$, are obtained by using recursive relationship (4), starting from (6). In practice, this is done by linear interpolation, separately for each draw of consumer-level parameters, θ_i. Then, the vector of reservation utilities is obtained by numerically solving a set of equations (8), also by linear interpolation.

□ **Click inequalities.** Suppose a consumer i has searched t_i pages and stopped. The page specific utilities are $u_{1i}, .., u_{t_i i}$ plus the outside option, u_{i0}. To avoid tracking u_{i0}, we will further include it in the utility of the first page (e.g., no click can be viewed as a 16th "hotel" on that page). Let $1 \leq k_i \leq t_i$ be index of the page clicked, which implies that this page delivers the highest utility:

$$u_{k_i i} \geq u_{mi}, \; m = 1 \ldots t. \tag{11}$$

Because the utility of the clicked page $u_{k_i i}$ is also the utility of the clicked hotel on that page, there is a set of inequalities related to other hotels that were displayed but not clicked on page k_i. Denote by $x_{k_i i}$ the utility of the clicked hotel (where $x_{k_i i} = u_{k_i i}$), and by $y_{k_i i}$ - best utility among other hotels on the same page. The click optimality implies,

$$x_{k_i i} > y_{k_i i}. \tag{12}$$

Collecting click-related inequalities from (11) and 12,

$$x_{k_i i} \geq u_{gi}, \; g = 1 \ldots k_i - 1 \tag{13}$$

$$x_{k_i i} > y_{k_i i} \tag{14}$$

$$x_{k_i i} > u_{g_i i}, \; g = k_i + 1 \ldots t_i. \tag{15}$$

[8] In our model, utility of every hotel has an i.i.d shock with EV Type 2 distribution, which is a continuous distribution with infinite support. This guarantees the full support of the distribution of search results (e.g., maximal utility on a page of hotels).

□ **Optimality conditions on joint searching and clicking decisions.** We now combine click-related and search-related inequalities derived previously, to arrive at a set of inequalities that must be jointly satisfied. One difficulty is that search inequalities were formulated in terms of period status quo, whereas click inequalities are stated in terms of product utilities. To proceed, the search inequalities need to be restated in the space of product utilities. Because clicking and searching decisions are not independent, some of the inequalities in the combined set will be redundant.

We propose a taxonomy of observations, as indexed by t - length of search and k - index of the clicked page. To save on notation, consumer specific index i is suppressed within this section.

Case 1. $k > 1, k <= t$.

Optimal search and stopping decisions imply inequalities concerning the status quo in each period: $u_1^* < r_1, .., u_{t-1}^* < r_{t-1}$ and $u_t^* > r_t$. From click inequalities, we conclude that once the preferred option was discovered on page k, all current and future values of status quo are equal to the utility of the preferred hotel, x_k. Therefore, search inequalities can be restated as:

$$u_1^* < r_1, .., u_{k-1}^* < r_{k-1} \tag{16}$$

$$x_k < r_k, .., x_k < r_{t-1}, k < t \tag{17}$$

$$x_k > r_t, \tag{18}$$

where the second set of inequalities may be empty (when $k = t$).

Utilities of pages observed before the best choice was found: u_g, $g = 1..k - 1$, are part of inequalities (16) but not (17) or (18). Because $u_g \leq u_g^*$, the inequalities $u_1^* < r_1, .., u_{t-1}^* < r_{t-1}$ imply that any u_g from this set must satisfy: $u_g < r_g, .., u_g < r_{t-1}$. These can be summarized as

$$u_g < \rho_g^k \equiv \min\{r_g, .., r_{k-1}\},\ g = 1..k - 1. \tag{19}$$

When $k < t$, inequalities (17) that are related to x_k can be summarized using a single statistic, denoted by ρ_k^t:

$$x_k < \rho_k^t \equiv \min\{r_k, .., r_{t-1}\}, k < t. \tag{20}$$

Utilities on pages $g = k + 1..t$, as well as utilities on the clicked page, y_k, are not involved in search decisions. The last inequality is the optimal stopping decision:

$$x_k > r_t. \tag{21}$$

Together, inequalities (19), (20), and (21) summarize all constraints that search decisions imply on the utilities of the observed hotels, for Case 1.

Case 2. $k = 1, t = 1$.

It is the simplest case, when no search occurred:

$$x_k > r_1. \tag{22}$$

Case 3. $k = 1, t > 1$.

This case corresponds to observations where consumer decided to search but went back to the first page. There are only (20) and 21:

$$x_k < \rho_k^t \equiv \min\{r_1, .., r_{t-1}\} \tag{23}$$

$$x_k > r_t. \tag{24}$$

TABLE 1 Inequality Conditions Implied by Searching and Clicking Decisions

Clicked Page	Observed Pages	Search	Click
$k = 1$	$t = 1$	$x_k > r_t$	$x_k > y_k$
$k = 1$	$t > 1$	$x_k < \min\{r_k, .., r_{t-1}\}$	$x_k > y_k$
		$x_k > r_t$	$x_k > u_g, g = k + 1..t$
$k > 1$	$t > 1$	$x_k < \min\{r_k, .., r_{t-1}\}, k < t$	$x_k > y_k$
		$x_k > r_t$	$x_k > u_g, g = 1..k - 1$
		$u_g < \min\{r_g, .., r_{k-1}\},\ g = 1..k - 1$	$x_k > u_g, g = k + 1..t$

TABLE 2 Composition of the Sample by Search Strategies

	Obs	%
[1] Consider only the first page	8300	34.6%
[2] Only flip default sorted results	3108	13.0%
[3] Sort by price and flip	1209	5.0%
[4] Sort by distance and flip	313	1.3%
Structural Sample	**12,930**	**54.0%**
[5] Filter by distance to city center, within 10 miles	269	1.1%
[6] Filter by distance to city center, within 2 miles	264	1.1%
[7] Filter by distance to city center, within 5 miles	411	1.7%
[8] Filter by landmark—Navy Pier	244	1.0%
[9] Filter by landmark—O'Hare airport	430	1.8%
[10] Reset landmark filters	209	0.9%
[11] Filter by neighborhood—Gold Coast	264	1.1%
[12] Filter by neighborhood—Loop	227	0.9%
[13] Filter by price—maximum 200	505	2.1%
[14] Filter by price—maximum 300	278	1.2%
[15] Filter by price—maximum 400	218	0.9%
[16] Next page	1232	5.1%
[17] Sort by distance to center	263	1.1%
[18] Sort by increasing price	1264	5.3%
[19] Sort by decreasing star rating	283	1.2%
Total	6361	26.5%
Estimation sample	**19,291**	**80.5%**
Rare search actions (dropped obs)	4668	19.5%
Original sample	23959	100.0%

Note also that as the best choice is observed before searching, there are no inequalities of type (19).

Table 1 collects all inequalities implied by the observed clicking and searching decisions, in the space of product utilities. In the web Appendix, we integrate out unobserved product-specific shocks to derive individual likelihoods of click and search decisions, conditional on the chosen search strategy.[9]

[9] According to this logic, it follows that the value of information contained in search decisions must be proportional to the size of search cost: the higher the search cost, the more informative search decisions are about consumer preferences. The above inequalities illustrate why *static demand estimates are inconsistent* if choice sets are generated by search. Such estimation includes both consumers who clicked on the first page, and those who clicked elsewhere. Each type of observation involves different sets of truncation regions for the utilities of observed products. As a result, the likelihood function of a static discrete–choice model is misspecified. To see the economic meaning of this, suppose there are only two pages, with one hotel per page: hotel A on page one and hotel B on page two. Consider consumers who searched and discovered B, but went back to buy A. The purchase decision implies that A > B. However, the search decision says the opposite: this consumer was willing to pay a positive search cost in order to explore hotels that are similar to B. In other words, we continue to infer that A > B, but there is an upper limit to the relative preference of A over B.

FIGURE 1

SEARCH INTENSITY IN THE DATA

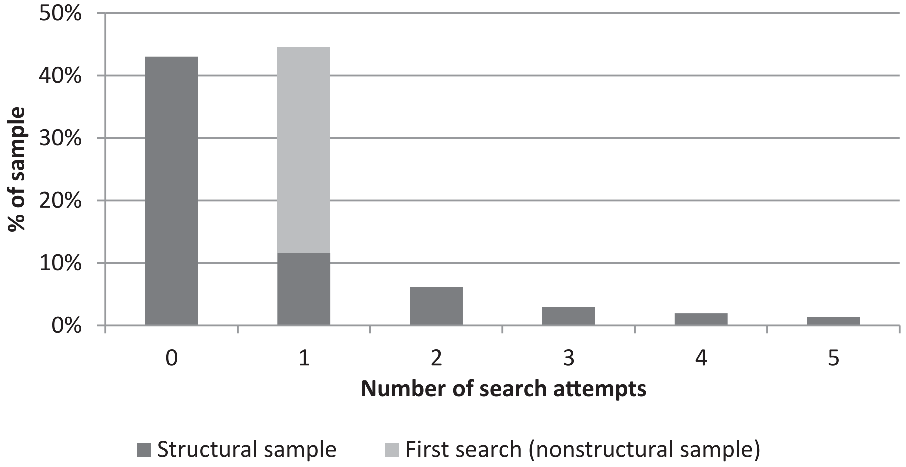


TABLE 3 Searching and Clicking Activity in the Structural Sample

N Pages Observed	% of Sample	Click Rate	% of All Clicks
1	64.2%	29.6%	61.9%
2	17.3%	34.3%	19.3%
3	9.1%	31.5%	9.4%
4	4.4%	33.3%	4.8%
5	2.9%	30.9%	2.9%
6	2.1%	26.8%	1.8%

Notes: Number of observations is 12,930 (structural sample).

TABLE 4 Nonprice Characteristics of Chicago Hotels in the Sample

Brand	Count	Neighborhood	Count	Stars	Count
None	34	Chinatown	3	One	9
Best Western	7	Gold Coast	51	Two	40
Hampton Inn	6	Loop	22	Three	55
Holiday Inn	6	South West	15	Four	42
Marriott	6	Midway	12	Five	2
Hilton	5	North Side	21		
Super 8	5	O'Hare	20		
Comfort Inn	4	West Side	3		
Hyatt	4				

Notes: In total, 148 Chicago hotels had online prices and were displayed to users in May 2007.

TABLE 5 Summary of Parameters of Request

	Minimum	Mean	Median	Maximum	Standard Deviation
Advance	1	33.50	21	364	36.63
Weekend	0	0.60	1	1	0.49
N days	1	2.44	2	30	1.65
N people	1	1.84	2	8	0.97

Notes: Number of observations is 23 959. Advance is the number of days between date of search and date of arrival. Weekend is a binary variation, equal to 1 if dates of stay include Saturday night.

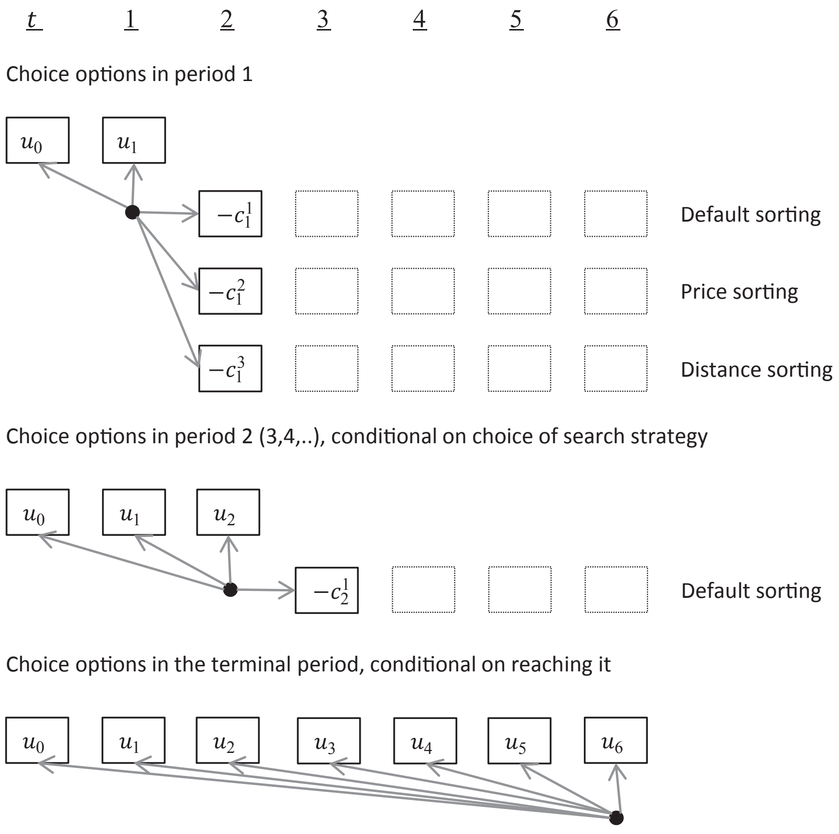
TABLE 6 Summary Statistics of Attributes of Hotels Displayed on the First Page

Var Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Price	3.1	2.4	2.7	2.4	2.3	2.5	2.5	2.4	2.3	2.3	2.3	2.3	2.4	2.4	2.4
	(1.6)	(1.2)	(1.3)	(1.2)	(1.1)	(1.1)	(1.2)	(1.2)	(1.2)	(1.3)	(1.3)	(1.3)	(1.3)	(1.3)	(1.3)
Star Rating	3.5	2.9	3.6	3.2	2.9	3.3	3.4	3.3	3.1	3.0	3.0	3.1	3.0	3.0	3.1
	(0.7)	(0.8)	(0.7)	(0.8)	(0.9)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.8)	(0.8)	(0.8)
Distance to	2.8	3.5	3.5	4.0	4.7	4.6	4.7	5.3	5.9	6.5	7.0	7.3	7.3	6.8	6.6
center	(4.8)	(5.2)	(5.6)	(5.8)	(5.9)	(6.2)	(6.1)	(6.3)	(6.2)	(6.2)	(6.3)	(6.4)	(6.5)	(6.6)	(6.5)
Distance to	13.3	13.0	12.6	12.6	12.2	12.5	12.2	12.1	12.0	11.7	11.5	11.1	11.4	11.7	11.3
O'Hare	(4.1)	(4.5)	(4.7)	(5.0)	(5.4)	(5.3)	(5.3)	(5.5)	(5.7)	(6.0)	(6.1)	(6.1)	(6.0)	(5.8)	(5.8)

Notes: Means and standard deviations (in brackets) of characteristics of hotels that occupied positions 1 to 15 on the first page. As every user observed a first page, the number of observations: 23,959.

FIGURE 2

CHOICE OPTIONS AT EVERY STAGE OF THE SEARCH PROCESS. IN BOXES SHOW IMMEDIATE PAYOFFS OF AN ACTION (PAY A SEARCH COST IF CONTINUE SEARCHING OR RECEIVE CLICK UTILITY). THE VALUE OF OUTSIDE OPTION IS u_0


t
1
2
3
4
5
6
Choice options in period 1
u_0
u_1
$-c_1^1$
Default sorting
$-c_1^2$
Price sorting
$-c_1^3$
Distance sorting
Choice options in period 2 (3,4,..), conditional on choice of search strategy
u_0
u_1
u_2
$-c_2^1$
Default sorting
Choice options in the terminal period, conditional on reaching it
u_0
u_1
u_2
u_3
u_4
u_5
u_6

The static choice model cannot capture this additional bit of information and therefore leads to biased inference.

A particular issue with search models is their ability to explain return, e.g., observations where the searcher returns and clicks on the previously found product. A recent article by de los Santos, Hortacsu, and Wildenbeest (2012) points out that a stationary sequential search model invariably predicts that a searcher will always buy the product found last. In contrast, in our data, about 20% of all clicks were made by consumers who searched but returned to the first page of results and clicked. The reconciliation between the theory and the data can be found in requirements on reservation utilities. Suppose $k = 1, t > 1$, which is Case 3 in our notation. For the set of inequalities (23)–(24) to be nonempty, reservation utilities must satisfy:

$$r_t < \min\{r_1, .., r_{t-1}\},\ \ k = 0, 1, t > 1. \tag{25}$$

Further, for the observations of Case 1 type, where $k > 1, k < t$, the following inequality must hold:

$$r_t < \min\{r_k, .., r_{t-1}\}, k > 1, k < t. \tag{26}$$

Observations of Case 2 type contain no return activity. Combining both cases, we conclude that declining reservation utilities, $r_1 > r_2 > r_3 ... > r_{T-1}$, is a sufficient, but not necessary, condition for rationality of return.

Potentially, there are two sources of nonstationarity of reservation utilities: (i) changing beliefs across pages; and (ii) increasing search costs. In our model, where beliefs are pinned by the empirical distribution of offers, the second mechanism is at work. A specification where every consumer is endowed with a single search cost is easily rejected by the data (the likelihood for a subset of observations are zero).

□ **Choice of search strategy.** After observing the first page of results, the consumer has a choice of search strategies by which to explore hotel options. The key observation is that the continuation value of a search strategy is simply its first-period reservation utility.

For strategy s and consumer i, let r^s_{1i} be the first-period reservation utility. Theoretically, consumers choose a strategy with the highest reservation utility. In practice, we find that the variation in reservation values is not sufficient to explain observed choices. A common approach to this problem is to add error term to each alternative:

$$V^s_i = r^s_{1i} + \varepsilon^s_i. \tag{27}$$

The alternative-specific shock ε^s_i be interpreted as a consumer's uncertainty about the value of search (e.g., future search cost) that is conducted according to the search strategy s.

The total number of alternative search strategies is $S = 3 + 15 = 18$. For $s = 1, 2, 3$, we compute continuation values in a fully structural way, that is, as a solution to the equation (8). These continuation values include both the potential benefit of immediate search results and the option value of continuing search. For strategies $s = 4..18$, the reservation utilities include only the immediate benefit of finding a better product. Concretely, the first-period reservation utility for $s = 4..18$ is found by solving $E^s_1 \max\{\tilde{u}_2 - r^s_{1i}, 0) = c_{1i}$. The index s indicates that the distribution of next-page utilities depends on the chosen search strategy. Once the reservation utilities are calculated, the unconditional likelihood is straightforward to obtain (see the web Appendix).

□ **Utility.** The displayed hotel information includes the name of the hotel, brand, price, geographical location, star rating, and amenities. Although more information on these hotels can be collected, we include only the displayed characteristics in our model. We also assume that, once the consumer observes the hotel's identity, she can costlessly infer her idiosyncratic taste about this hotel, or match value.[10] We choose the following specification:

$$u(p_j, q_j, \varepsilon_{ij}) = \alpha^p_i P_{ij} + \alpha_1 r_{ij} + \alpha_{2i} do_j + \alpha_{3i} d_j + \alpha_{4i} s_j + \vec{\alpha}_n \vec{n}_j + \vec{\alpha}_b \vec{b}_j + \varepsilon_{ij}, \tag{28}$$

where P_{ij} is the price of hotel j (in hundreds of dollars), displayed to consumer i; $q_j = (do_j, d_j, s_j, \vec{n}_j, \vec{b}_j)$ is a vector of nonprice characteristics of hotel j: distance to O'Hare

[10] Learning the match value can be costly. This cost can be modelled explicitly, as in Kim, Albuquerque, and Bronnenberg (2010), or implicitly, as in this article, where it constitutes a part of the search cost.

airport, distance to the city center, star rating, and a set of neighborhood and chain dummies. We take $d_j = \log(1 + D_j)$ —the logarithm of distance (in miles), in order to smooth the outliers. Both distance metrics—to city center, and to O'Hare airport—are included in the model. Because hotels are not located on a straight line, these metrics are not collinear and represent important attributes of demand for hotels.[11]

Differences in quality standards between hotel chains are captured by a set of chain dummies, $\overrightarrow{b}_j$. A large number of hotel brands are present in the Chicago market, but for most of them only a few clicks are observed. Therefore, we include only the most popular brands, shown in Table 4, which together attract 28% of impressions and 56% of clicks. The ”none” option stands for independent hotels; all other hotels are grouped under a default category.

In the data, there is strong correlation between the position of the hotel on the page and its click rate. To explain this phenomenon, we introduce r_{ij}—hotel position on the page—as a separate attribute, which captures in a reduced form way the intrapage search that we otherwise do not observe.

Consumer tastes for price and nonprice characteristics are allowed to vary with parameters of request. The information in the search request is summarized by a vector of four dummies, denoted by R_i, which includes (i) whether the search is made more than 30 days in advance, (ii) whether stay includes two nights or longer, (iii) whether Saturday night stay is included, (iv) whether the person is traveling alone. Further, we introduce unobserved variation in tastes, mainly through the parameter of price sensitivity, α_i^p.

Leaving the website without any click constitutes a choice of the outside option, whose mean utility also may depend on consumer-level characteristics:

$$u_{i0} = \mu_0 + \mu_1 R_i + \varepsilon_{i0}. \tag{29}$$

By parameters of request R_i in the value of the outside option, we attempt to control for various reasons for why a user may leave the website. For example, the user may decide to call the hotel directly, or to search later, or to abandon the idea of the trip. Even though we do not observe all these reasons, we may conjecture that users who search farther in advance have more opportunities for searching later and hence are less likely to settle at the moment. Note that the utility specification (28) does not include a constant term, which is a necessary exclusion restriction to identify μ_0.

□ **Beliefs.** To determine the expected benefit of search, the consumer formulates a belief about distribution of search results: the maximal utility of hotels on the next page. It is constructed in two steps. We first specify the distribution of hotel characteristics on the next page: prices, qualities, and match values. Then, we use the utility model to map beliefs from the multidimensional space of product characteristics into the single dimension of utilities.[12]

Let $G_t^s(p_j, q_j, \varepsilon_{ij})$ be the consumer's belief about the joint distribution of attributes of a random hotel on the next page, if the search continues according to strategy s. With the rest of the literature, we assume that G_t^s reflects the actual distribution offers, that is, we estimate search from known distribution. Using the chain rule and the independence of taste shocks, we can rewrite G_t^s as a product of conditionals:

$$G_t^s(p_j, q_j, \varepsilon_{ij}) = H_t^s(p_j, q_j) f_\varepsilon(\varepsilon_{ij}), \tag{30}$$

[11] It is possible that searchers who want to stay close to the airport care only about distance to O'Hare, but not about distance to the city center, and vice versa. To some extent, we capture these differences by including interaction terms between parameters of request and distance to the city center (advance purchase, weekend stay, and number of travelers).

[12] Assumptions on consumer beliefs regarding search are central to any search model. Search cost estimates are generally quite sensitive to the location and scale parameters of the distribution of search results. Therefore, it is important that our approach to construction of beliefs is empirically driven by distributions of actual search results found in the data. Thanks to an anonymous referee for this point.

where the distribution of match values, $f_\varepsilon(\varepsilon_{ij})$, is assumed to be EV Type 1, as in the utility model. That is, consumers do not know the realizations of their tastes for hotels that will show up on the next page; another way to put it is that consumers do not know the identities of hotels that will be discovered.

The joint distribution of observable hotel characteristics on page t, $H^s_t(p_j, q_j)$, is approximated by taking bootstrap samples from the actual contents of page t as seen by consumers who searched using strategy s. On a given simulation draw r, we generate a vector $\{(p^r_1, q^r_1), .., (p^r_{15}, q^r_{15})\}$ of 15 hotels and their prices that may appear on the next page as a result of search. The maximal utility among these hotels is EV Type 1 with scale one and location parameter $M^r_t = \log(\exp(\mu(p^r_1, q^r_1)) + .. + \exp(\mu(p^r_{15}, q^r_{15})))$. Therefore, we simulate maximal utility as $u^r_t = M^r_t + \varepsilon^r_t$. Repeating the process for $r = 1..R$ times, we obtain an $R \times 1$ vector of random draws of maximal utilities on the next page.

4. Identification

■ In principle, the full identification of a search model requires that we are able to uniquely recover the joint distribution of preferences, search costs, and beliefs. We make several assumptions that simplify this problem. First, we assume that consumer beliefs can be reasonably approximated by the empirical distribution of hotel prices and qualities. This is a standard assumption in the empirical literature on consumer search (see, e.g., Hong and Shum, 2006), which is grounded in the idea that beliefs should be rational. Second, we assume a common mean utility function: $u_{ij} = \mu(p_j, q_j) + \varepsilon_{ij}$. Under these assumptions, we show that the mean utility function and the distribution of search costs are nonparametrically identified.

Even though the assumption of logit demand is restrictive, our result represents the first step toward establishing the identification of a search model with differentiated products. Existing results assume either search for best price (no differentiation), or vertical differentiation, neither of which can rationalize consumer choices observed in our data set.

□ Mean utility function.

We start with the *identification of mean utility function, $\mu(p, q)$*. Consider a population of consumers who entered the same request R and observed the same contents of the first page. Let P_h be the proportion of those who clicked on the first page, on a hotel h with characteristics (p_h, q_h). These include consumers who clicked without searching, and those who searched further but returned to the first page.[13] Also, let P_0 be the proportion of those who chose outside option, whose mean utility is $\mu_0(R)$. Because P_h and P_0 are observed in our data, we can compute their ratio, $P_{h0} = P_h/P_0$. The search model predicts the following relationship to mean utility functions: $P_{h0} = \exp(\mu(p_h, q_h) - \mu_0(R))$. By inverting this equation, we obtain the value of the function $\delta(p, q, R) = \mu(p, q) - \mu_0(R)$ at the point (p_h, q_h, R). The function $\mu_0(R)$ is identified from $-\delta(p_0, q_0, R)$ up to a constant, $c = -\mu(p_0, q_0)$; conversely, the function $\mu(p, q)$ is identified from $\delta(p, q, R_0) - \delta(p_0, q_0, R_0)$ up to the same constant.

□ Uncovering the search cost distribution.

Turning to the *identification of the search cost distribution*, consider a population of users who observed the same first page content, denoted by $\Omega_1 = \{p_r, q_r\}^{15}_{r=1}$ - prices and qualities of hotels. From our data, we can compute the share of searchers who observed the same first page: $P(T > 1|\Omega_1)$. Applying the search rule (9) and integrating over search cost, we obtain the predicted share of searchers:

$$P(T > 1|\Omega_1) = \int_0^{+\infty} P(u_{i0} < r_1(c), .., u_{i15} < r_1(c))g(c)dc. \tag{31}$$

[13] Note that we did not use observations for consumers who clicked beyond the first page, because their likelihood contributions do not have the necessary multiplicative forms. Please consult the web Appendix for analytical results.

Using the definition of extreme value distribution,

$$P(u_{i0} < r_1(c), .., u_{i15} < r_1(c)) = \prod_{r=0}^{15} \exp(-\exp(-r_1(c) + \mu(p_r, q_r)))$$

$$= \exp\left(-\sum_{r=0}^{15} \exp(-r_1(c) + \mu_r)\right)$$

$$= \exp\left(-\exp(-r_1(c)) \sum_{r=0}^{15} \exp(\mu_r)\right).$$

In this formula, $S(\Omega_1) = \sum_{r=0}^{15} \exp(\mu_r)$ is a sufficient statistic that summarizes the relationship between the content of the first page, Ω_1, and the search decision. Because mean utilities are already identified, this function is also known. We can rewrite the search decision as,

$$P(T > 1|\Omega_1) = \int_0^{+\infty} \exp(-\exp(-r_1(c))S(\Omega_1))\, g(c)dc. \tag{32}$$

With $r_1(c)$ being a monotonic function, we can introduce a change of variables, $t = \exp(-r_1(c))$. The above equation becomes,

$$P(S) = \int_0^{+\infty} \exp(-tS)\, h(t)dt, \tag{33}$$

where $h(t) = g(c^{-1}(t))/t\prime(c)$. With respect to unknown density $h(t)$, equation (33) is a Friedholm Type 1 integral equation.[14] As the kernel of this integral equation belongs to the exponential family, the solution is unique (Lehmann and Romano, 2005). As usual, the existence is guaranteed by the assumption that the model is correct.

The identification strategy outlined above rests on two assumptions: that a hotel's price, as observed by a consumer, is uncorrelated with a consumer's idiosyncratic tastes and that a hotel's first-page membership is also uncorrelated with tastes. Although we cannot guarantee that these assumptions hold in our setup, there are factors that alleviate potential endogeneity concerns. See the web Appendix for a detailed discussion.

5. External evidence on search costs

■ Before considering the estimates of search costs delivered by the search model, it is helpful to obtain an external estimate of the possible magnitude of search costs. Consider a consumer who decided not to search. Her search cost satisfies the following inequality:

$$c_i \geq \bar{c}_i = E \max\{\tilde{u}_2 - u_{1i}, 0\},$$

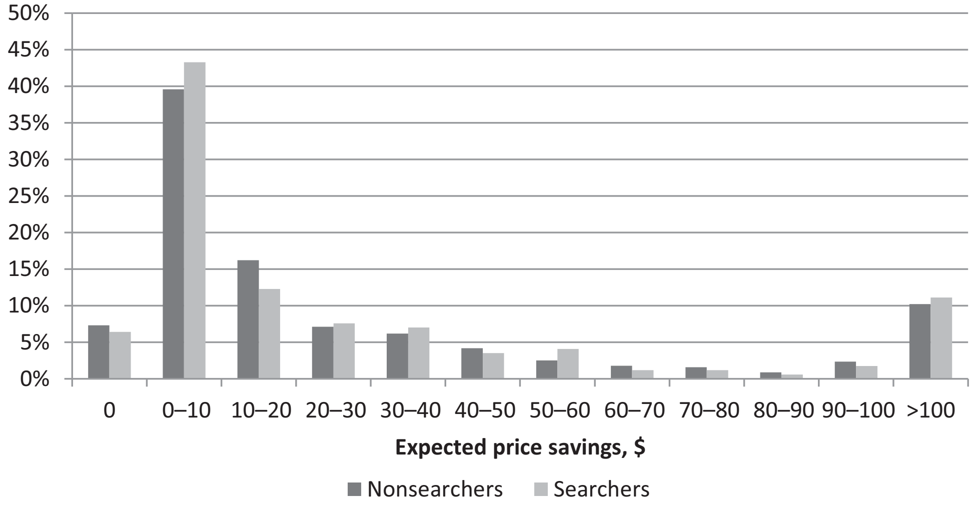
where u_{1i}—the status quo (best utility on the first page), and $\tilde{u}_2$—unobserved (by consumer) utility of the best hotel on the second page of results. (In fact, the actual threshold is higher than $\bar{c}_i$ because of a positive option value of searching further). By computing the distribution of search cost cutoffs $\bar{c}_i$ among consumers who did not search, we obtain the distribution of lower bounds on search costs. Similarly, a distribution of upper bounds on search costs is obtained from consumers who did search.

Search cost cutoffs $\bar{c}_i$ cannot be computed directly, because the consumer-specific status quo, u_{1i}, is unobserved. However, an approximation can be obtained from a subsample of consumers who actually clicked on a first page hotel, revealing the price and nonprice characteristics of their

[14] In our data, the variation in the content of the first page is so frequent that the ”market” typically consists of only one consumer. Because of this, conditional probabilities like (32) cannot be inverted. Therefore, only a parametric version of the model is feasible to estimate. Even so, the variation of prices of hotels on the first page provides a wealth of identifying equations of the type (31) that allow for a flexible parametric form assumptions.

FIGURE 3

EXPECTED SEARCH BENEFIT AMONG CONSUMERS WHO CLICKED ON THE FIRST PAGE



status quo option. In our data, there are 2458 consumers who clicked without searching, and 171 consumers who searched but returned and clicked on a first page hotel.

For each consumer, we record the star rating and neighborhood of the clicked hotel, as well as the hotel's price. Conservatively, we assume that the consumer is interested only in hotels with the same combination of star rating and neighborhood as the clicked hotel. Then, the benefit of search is computed as expected price savings among hotels in the preferred category, among 15 hotels on the second page.

Figure 3 depicts the distribution of expected price savings, for searchers and nonsearchers, both curves looking similar. For about 7% of observations, we do not find any potential price savings. For 56% of the sample, the expected savings are \$20 or less (the median search cost cutoff is \$11.3). Finally, there is a sizeable portion of consumers with expected benefits exceeding \$50 (19% among nonsearchers, and 20% among searchers).

6. Estimation results

■ The primary goal of the empirical implementation of the search model is to obtain estimates of search costs in this environment. This, in turn, allows to compute various quantities of interest: (i) price elasticities of search-generated demand and (ii) outcomes of counterfactual policies.[15]

We estimate several specifications, to investigate the sensitivity of search cost estimates to our assumptions on utility and search cost distribution. The estimation is done by Simulated Maximum Likelihood, numerically integrating individual likelihood of joint clicking and searching decisions, over unobserved search costs and random tastes. In practice, we take 200 Halton draws from the distribution of search costs, as well as any random tastes if included in the model. The standard errors are adjusted upward to include the effect of simulation noise, using methods discussed in Train (2003).

□ **Preferences.** Table 7 presents the first set of results—estimates of preference parameters. Columns D2 and D2R present estimates from a multinomial logit model of click, conditional

[15] See the web Appendix for two examples of counterfactual policies. See also a companion article, de los Santos and Koulayev (2013), where we use a search model to construct a recommendation ranking on a search platform.

TABLE 7 Estimates of Utility Parameters under Different Specifications

	D2		DR2		S2		S2R	
Price	−1.18	(0.03)			−0.96	(0.03)		
Star rating	0.50	(0.02)	0.65	(0.02)	0.38	(0.02)	0.30	(0.02)
Distance to center	−0.49	(0.04)	−0.59	(0.05)	−1.09	(0.04)	−1.05	(0.04)
Distance to O'Hare	0.23	(0.07)	0.86	(0.07)	0.23	(0.07)	0.26	(0.00)
Position on the page	−0.10	(0.00)	−0.07	(0.00)	−0.12	(0.00)		
Price coef median			−2.51	(0.09)			−0.80	(0.02)
Price coef SD			1.66	(0.67)			0.26	(0.04)
Interactions of Hotel Characteristics and Request Parameters								
Intercepts								
Advance <=30 days	0.60	(0.06)	0.26	(0.08)	−0.15	(0.04)	−0.03	(0.05)
Number of days <=2	−0.17	(0.05)	−0.08	(0.07)	0.01	(0.03)	−0.03	(0.03)
Weekend stay	0.75	(0.06)	0.54	(0.08)	0.06	(0.04)	0.16	(0.05)
Traveling alone	0.30	(0.05)	0.50	(0.07)	0.11	(0.03)	0.20	(0.04)
Price interactions								
(Advance <=30) × Price	0.44	(0.03)	0.11	(0.04)	0.07	(0.02)	0.13	(0.02)
(Number of days <=2) × Price	−0.15	(0.02)	−0.22	(0.03)	0.07	(0.01)	−0.07	(0.01)
(Weekend stay) × Price	0.32	(0.03)	0.24	(0.03)	0.16	(0.02)	0.12	(0.02)
(Traveling alone) × Price	0.06	(0.02)	0.00	(0.03)	0.10	(0.02)	0.01	(0.02)
Quality interactions								
(Advance <=30) × (Stars <=2)	0.33	(0.05)	0.12	(0.05)	0.09	(0.04)	0.17	(0.05)
(Weekend stay) × (Stars <=2)	0.55	(0.05)	0.39	(0.06)	0.13	(0.04)	0.18	(0.05)
(Advance <=30) × (Within 1 mile from center)	−0.53	(0.06)	−0.39	(0.06)	−0.50	(0.04)	−0.46	(0.05)
(Traveling alone) × (Within 1 mile from center)	−0.15	(0.05)	−0.25	(0.06)	−0.33	(0.04)	−0.23	(0.05)
(Weekend stay) × (Within 1 mile from center)	0.02	(0.05)	−0.01	(0.05)	−0.16	(0.04)	−0.19	(0.04)
Derived Variables of Interest								
Position value($)	8.51	(0.32)	2.93	(0.15)	12.97	(0.50)	18.78	(0.52)
WTP for star rating	42.66	(2.27)	25.81	(1.36)	39.56	(2.49)	37.46	(2.61)
Price elasticity	−1.62	(0.15)	−1.73	(0.10)	−1.27	(0.05)	−0.93	(0.10)
Log-likelihood (1000s)	38.75		37.62		78.05		77.68	

Notes: Estimates of preferences for hotels, from static and dynamic (search) models. D2—constant coefficient static model with actual choice sets. Dependent variable: click conditional on actual choice set. D2R—same as above, but with random coefficients. S2—constant coefficient search model with log-normal baseline search cost. Dependent variable: joint search and click. S2R—random coefficient search model with log-normal baseline cost. In all models, the number of observations is 19,291.

on the observed choice sets.[16] Model D2 assumes that preference parameters conditional on consumer observables are constant. Model D2R relaxes this assumption by introducing random price sensitivity parameter α_i with negative log-normal distribution. Estimates show substantial variation in the unobserved component of price sensitivity, which is not surprising, given that such consumer characteristics as income, age, etc., are not observed.

Static discrete-choice models such as D2 and D2R are useful as a starting point of the analysis. These are widely used models and require only click data. However, their estimates are biased, and it is of interest to explore the magnitude of the bias. Because one cannot directly compare estimates of nonlinear models, we construct three derived quantities: (i) dollar value of hotel's position, which is a change in hotel price that is equivalent, in terms of expected click rate, to a change in hotel's display rank by one unit; (ii) dollar value of one unit of star rating; (iii) own price elasticity of demand, as a percentage change in expected click rate following 1% increase in price.

[16] Formally, for each consumer we observe choice set C_i and a click on a hotel $h_i \in C_i$. The likelihood is:

$$P_i(h_i) = \int \frac{\exp(\alpha_i p_{ih_i} + \mu_i(X_{h_i}))}{1 + \sum_j \exp(\alpha_i p_{ij} + \mu_i(X_j))} dF(\alpha_i),$$

where p_{ih_i} - price of hotel h_i as observed by consumer i, and $\mu_i(X_{h_i})$ - nonprice part of the mean utility.

Columns S2 and S2R in Table 7 present estimates from search models, whose assumptions on preferences mirror those of D2 and D2R. In S2, tastes are conditionally constant, and in S2R, we include unobserved heterogeneity in price coefficient, which allows for somewhat more flexible substitution patterns.[17] Note that contrary to D2, the search-generated demand in model S2 does not have IIA property, because of the unobserved heterogeneity in search costs.

Overall, the estimates of hotel preferences have economically meaningful signs and magnitudes, which is a particularly good sign, given that our data comes from clicks, not purchases. We find that demand for hotel clicks is relative elastic, well above one for all models. When computing elasticity, we consider clicks for a particular hotel: located 1.13 miles from center, three stars, \$235 average price. As a counterfactual, this hotel is located on the first page, first position, for all consumers in the sample. Estimates of price elasticity from click-only models range between -1.62 for D2 to -1.73 for D2R. Search models S2 and S2R give values of -1.27 and -0.93, respectively. Differences in the predicted value of position or the value of star rating are also substantial.

Results from Table 7 suggest that request parameters are informative about consumer preferences: most of the interactions of weekend stay, advance purchase, and number of travelers with a hotel's attributes are statistically and economically significant. Even though our primary interest is in the estimates of search costs, we find that a parsimonious specification of preferences is important. If consumer A is primarily interested in cheaper hotels, located further away from city center, then the potential presence of more expensive hotels among search results does not improve the benefit of search. Vice versa, if consumer B is price insensitive and would like to stay closer to city center, the possibility of finding airport hotels will not motivate her to search further. If a search model cannot distinguish between two types, it will overstate the variance of search results.

□ **Search costs.** At first, it may appear that search costs should be small: another batch of results is only a click away. However, there is evidence that consumers often forgo material price savings by not searching. We interpret search cost as cognitive costs of processing information on a new page of results: indeed, comparing prices and various quality characteristics of 15 new hotels is not a trivial task. Note that in our model, the search cost implicitly includes the cost of learning the match quality.[18]

In all search specifications we estimate, search costs are allowed to change as the search progresses. As shown in equation (2), a consumer is characterized by an idiosyncratic baseline search cost, which is the cost of making the first search attempt. The cost of each subsequent search attempt is equal to the baseline plus a positive increment. Increments are constant across consumers, but independent across time periods. This latter assumption is motivated by our data, where the search intensity decreases in an uneven fashion, as seen from Figure 1. The dynamics of search intensity identifies the increments, and the average level of search identifies the baseline cost in our model.

Table 8 presents main results regarding search costs. Dollar values are reported. As discussed previously, we explicitly model all search decisions made by consumers who chose the top three search strategies, and only the first search action by consumers who chose other strategies (see Table 2). Generally, search costs with vary by search strategy, and we find this feature of the

[17] Although not a formal proof, the intuition behind the identification of search costs with random tastes is this. Both random tastes and search costs affect observed search decisions: for instance, more price sensitive consumers are more likely to search, as for them the variation in search results is larger. Alternatively, a lower search cost can also lead to more search. However, only random tastes affect the distribution of clicks conditional on choice sets, once the search is over. That is, with enough variation in clicks and choice sets, we have enough data to identify random tastes (taking into account search inequalities, of course). Given that, search costs are identified from search decisions. In addition, when products have more attributes than just price, there is no longer clear substitution between search costs and tastes in the effect on search decisions.

[18] We are grateful to an anonymous referee for this observation.

TABLE 8 Search Cost Estimates under Various Specifications

Model	S1	S2	S2R	S3	–	–
Medians of Search Costs for Top three Strategies, by Search Attempt						
	Top three Strategies	Top three Strategies	Top three Strategies	Top one — Default Sorting	Top two — Price Sorting	Top three — Distance Sorting
First search	3.79	3.04	4.06	1.59	3.71	5.17
	(0.23)	(0.19)	(0.19)	(0.09)	(0.20)	(0.51)
Second search	10.51	9.33	11.56	6.56	7.82	5.76
	(0.43)	(0.39)	(0.40)	(0.23)	(0.37)	(1.70)
Third search	8.58	8.04	10.34	5.64	6.11	5.21
	(0.40)	(0.37)	(0.43)	(0.24)	(2.87)	(0.51)
Fourth search	13.64	11.71	14.70	8.98	7.66	8.63
	(0.52)	(0.49)	(0.54)	(0.33)	(3.84)	(1.26)
Fifth search	15.60	12.56	16.38	11.06	5.13	
	(0.66)	(0.59)	(0.72)	(0.46)	(3.98)	
Standard Deviation of Search Cost						
First search		5.50	9.88	3.09	7.22	10.07
		(0.55)	(1.22)	(0.47)	(1.32)	(1.83)
Interactions of Baseline Search Cost with Parameters of Request						
Advance <=30 days	0.27	0.18	−0.02	0.20		
	(0.04)	(0.04)	(0.03)	(0.04)		
Weekend stay	0.32	0.22	−0.02	0.42		
	(0.04)	(0.04)	(0.03)	(0.04)		
Traveling alone	−0.16	0.01	−0.30	0.06		
	(0.04)	(0.04)	(0.03)	(0.04)		
Medians of Search Costs for Other Strategies						
First search	8.68	5.83	3.59	4.71		
	(0.51)	(0.37)	(0.21)	(0.26)		

Notes: Shown are dollar estimates of search costs for models: S1—constant coefficient model with constant search costs; S2—constant coefficient model with log-normal baseline search cost; S2R—random coefficient model with log-normal baseline cost; S3—constant coefficient model with log-normal baseline, search costs may vary by strategy.

model to be empirically important. Models S1, S2, S2R in Table 8 assume equal search costs across the top three strategies, and a different cost for other strategies. Model S3 is more flexible in that it allows consumers to have a different search cost when they search by price sorting than if they search by flipping recommended hotels (default sorting). In models S2, S2R, and S3, the baseline search cost has log-normal distribution (standard deviation is reported at the bottom of the table), in S1, search costs are constant.

Figure 4 plots the evolution of search costs by each of the top three search strategies (exact values and standard errors are found in the last three columns of Table 8). We find that search costs increase with each search attempt. The highest gradient is found with default sorting strategy, where consumers search by flipping pages of recommended hotels: median search cost increases from under $2 to over $12. A possible reason is that consumers actually do not observe the underlying ranking criterion (as opposed to, say sorting by price or distance to city center) so that they become discouraged more quickly. It is notable that standard deviations of baseline search costs are large, across all models with log-normal distribution (S2, S2R, S3).

Although a log-normal distribution is a natural modelling choice for search costs—everywhere positive, two-parameter distribution—by being unimodal, it may miss important features of search cost heterogeneity. Therefore, we experiment with more flexible forms of distribution for the baseline search cost. Figure 5 compares log-normal distribution (model S2) to a mixture of two log-normals.[19] We find mixture components to be well identified, one with lower

[19] In our results, estimates of utility parameters are not dramatically affected by switching to the mixture of log-normals (as compared to model S2, with a log-normal search cost). At the same time, the estimate of price elasticity

FIGURE 4

MEDIAN SEARCH COST, BY SEARCH STRATEGY AND SEARCH ATTEMPT

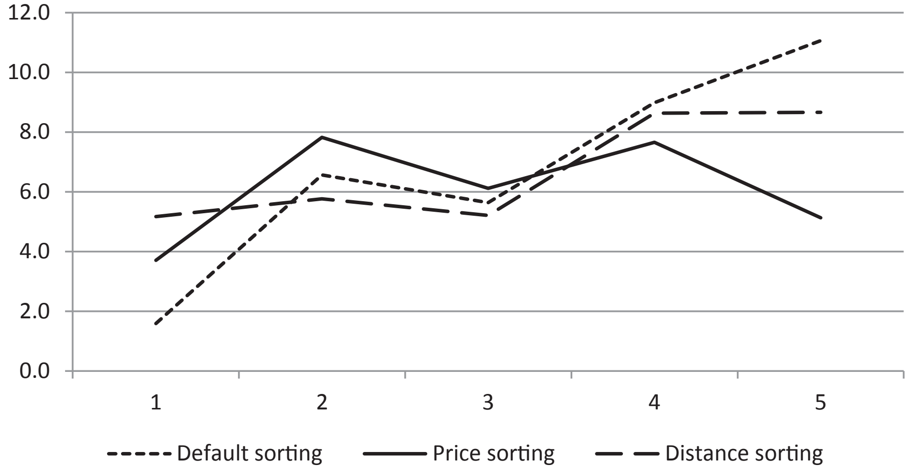
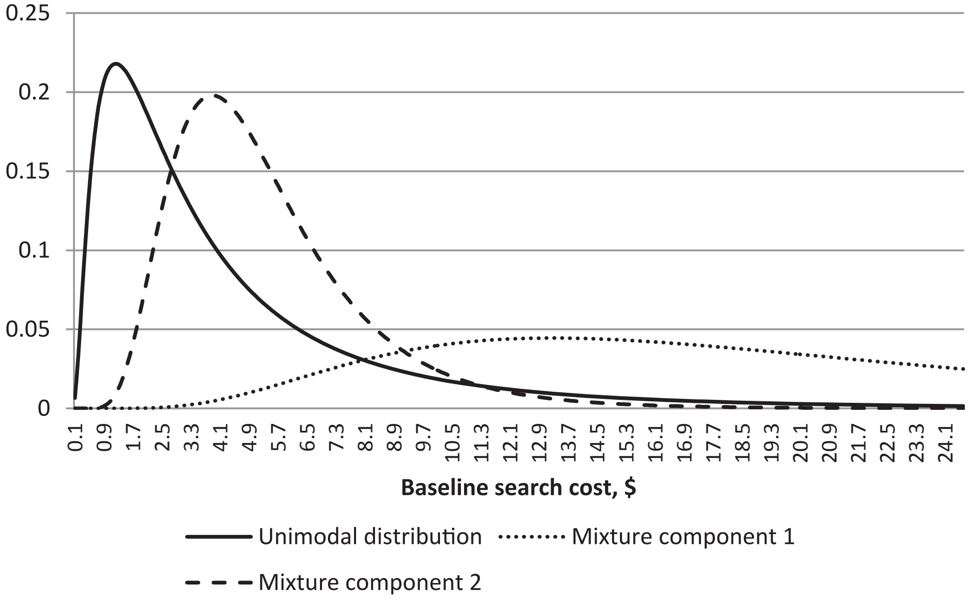


FIGURE 5

LOG-NORMAL SEARCH COST DISTRIBUTIONS

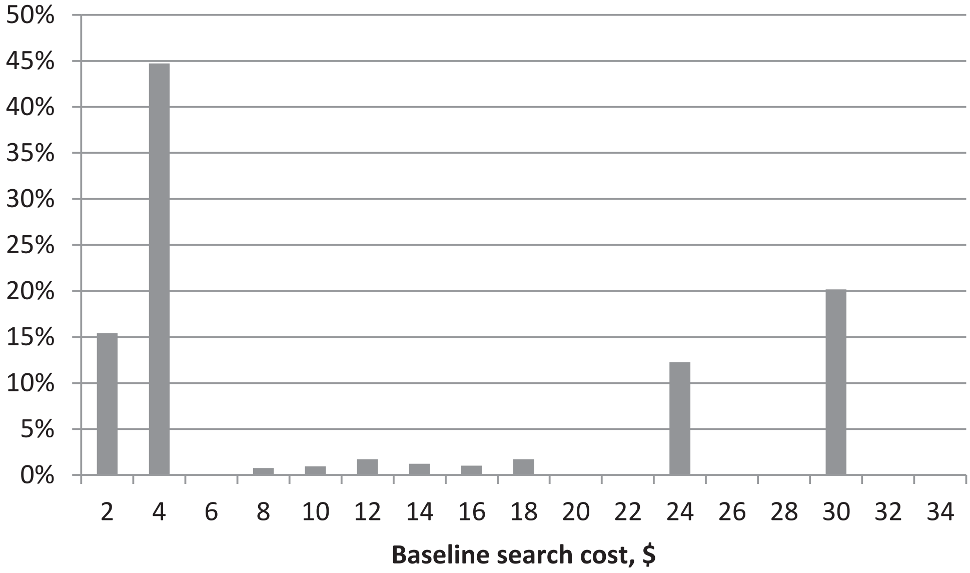


median of \$3.6 and population weight of 54%, and another with median \$12.4, with weight 46%. It is also notable that the variance of the second, higher cost type is also much larger. The issue of multimodality of search costs is further illustrated in Figure 6 that depicts a discrete distribution of baseline search costs, with freely estimated weights. With more flexible specification, the multimodality becomes more apparent. About 20% of consumers have search costs as large as \$30, and another 12% with search cost of \$24. These magnitudes are necessary to rationalize

changes from −1.27 in S2 to −1.88 with mixture. This is explained by the fact that in addition to utility parameters, search costs also play an important role in in the price elasticity.

FIGURE 6

DISCRETE SEARCH COST DISTRIBUTION



consumer behavior where a nontrivial proportion of consumers forego large price savings by not searching.

□ **Estimates conditional on a search strategy.** We additionally estimate search models using subsamples of consumers who chose a particular search strategy. We choose the two most popular strategies: (i) those who searched by flipping pages of recommended hotels and (ii) those who searched by price sorting. Because all these consumers made at least one search attempt, we can only estimate the cost of their second, third, etc., search attempt. Results are presented in Table 9.

Comparing the estimates of preferences to those obtained under the choice of strategy model (column S2 in Table 7), we can see that both subsamples are highly selected. For instance, consumers who chose to search by price sorting are naturally found to be more price sensitive. This is not surprising, as the choice of how to search is essentially a preference over a subset of products to be explored. Perhaps more interesting are implications for estimates of search costs. *A priori*, there are several additional identifying restrictions that a choice of strategy imposes on search cost:

First, dynamic effects of future search costs on current search decisions: when a consumer makes a search decision in the first period, one of the benefits of search is an option value of continuing to search later; however, if later search is more costly, that option value is reduced.

Second, even though search costs may differ by search strategy, they are not independently distributed. Therefore, observations where consumers choose, say, price sorting, will have an impact on the search cost estimates for other strategies.

Third, the decision not to search is affected by the variety of available search strategies; the larger is the variety of choice, the higher search costs are needed to rationalize nonactivity.

All these relationships are not captured in the conditional search model, which misses the important first step where consumers decide how to search. Results in Table 9 can be compared to model S3 in Table 8, where search cost medians are allowed to vary by strategy. We find that for price sorters, the conditional search model somewhat overestimates median search costs, as compared to the full model. The result is the opposite for those who searched recommended hotels.

TABLE 9 Estimates Conditional on a Search Strategy

	Search Recommended Hotels		Search by Price Sorting	
Price	−1.28	(0.15)	−1.61	(0.34)
Star rating	0.38	(0.06)	0.32	(0.10)
Distance to center	−0.39	(0.11)	−1.21	(0.22)
Distance to O'Hare	0.27	(0.20)	−0.09	(0.23)
Position on the page	−0.03	(0.01)	−0.14	(0.02)
Interactions of Hotel Characteristics and Request Parameters				
	Yes		Yes	
Derived Variables of Interest				
Position value($)	2.41	(0.46)	8.44	(2.53)
WTP for star rating	30.03	(4.18)	20.00	(8.59)
Price elasticity	−1.85	(0.00)	−4.81	(0.00)
Medians of Search Costs, by Search Attempt				
First search	N/A		N/A	
Second search	3.25	(0.25)	8.16	(1.33)
Third search	6.53	(0.41)	14.55	(0.89)
Fourth search	4.14	(0.29)	11.52	(1.00)
Fifth search	7.51	(0.49)	10.27	(1.13)
Standard Deviation of the Baseline Search Cost				
Second search	0.64	(0.97)	5.78	(1.03)
Interactions of Baseline Search Cost with Parameters of Request				
	Yes		No	
number of observations	3108		1209	

Notes: Estimates from search models conditional on the choice of search strategy. The first column—estimates on a subsample of consumers who searched by flipping recommended hotels. The second column—subsample of consumers who searched by price sorting (and then flipping sorted results).

7. Conclusions

■ In this article, we have estimated a structural model of consumer search for differentiated products, using a unique data set of search histories by consumers looking to book a hotel online. The estimation implements a novel identification strategy that we propose, in order to separate the impact of unobserved heterogeneous search costs and preferences on search decisions. We find that search costs are substantial, and therefore, search frictions play an important role in shaping consumer demand. The approach and methods we develop are applicable in a wide variety of situations, where consumers search before making a purchase.

References

BERRY, S. "Estimating Discrete-Choice Models of Product Differentiation." *The RAND Journal of Economics*, Vol. 25(2) (1994), pp. 242–262.

BERRY, S., LEVINSOHN, J. AND PAKES, A. . "Automobile Prices in Market Equilibrium." *Econometrica*, Vol. 63(4) (1995), pp. 841–890.

BRYNJOLFSSON, E., DICK, A. AND SMITH, M. "A Nearly Perfect Market? Differentiation Versus. Price in Consumer Choice." *Quantitative Marketing and Economics*, Vol. 8(1) (2010), pp. 1–33.

BRYNJOLFSSON, E. AND SMITH, M. "Consumer Decision-Making at an Internet Shopbot: Brand Still Matters." *The Journal of Industrial Economics*, Vol. 49(4) (2001), pp. 541–558.

DE LOS SANTOS, B. "Consumer Search on the Internet." NET Institute Working Paper no. 08–15, 2008.

DE LOS SANTOS, B., HORTACSU, A. AND WILDENBEEST, M. "Testing Models of Consumer Search Using Data on Web Browsing and Purchasing Behavior." *American Economic Review*, Vol. 102(6) (2012), pp. 2955–2980.

DE LOS SANTOS, B. AND KOULAYEV, S. "Optimizing Click-Through in Online Rankings for Partially Anonymous Consumers." Unpublished manuscript, 2013.

GHOSE, A., IPEIROTIS, P., AND LI, B. "Surviving Social Media Overload: Predicting Consumer Footprints on Product Search Engines." Unpublished manuscript, 2012.

Hong, H. and Shum, M. (2006), "Using Price Distributions to Estimate Search Costs." *RAND Journal of Economics*, Vol. 37(2) (2006), pp. 257–275.

Hortacsu, A. and Syverson, C. "Product Differentiation, Search Costs, and Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds." *Quarterly Journal of Economics*, Vol. 119 (May 2004), pp. 403–456.

Johnson, E. J., Moe, W. W, Fader, P. S., Bellman, S., and Lohse, G. L. (2004). "On the Depth and Dynamics of Online Search Behavior." *Management Science*, Vol. 50(3) 299–308.

Kim, J. B., Albuquerque, P. and Bronnenberg, B.J. "Online Demand under Limited Consumer Search." *Marketing Science*, Vol. 29 (6) (2010), pp. 1101–1023.

Koulayev, S. "Search with Dirichlet Priors: Estimation and Implications for Consumer Demand." *Journal of Business and Economic Statistics*, Vol. 31(2) (2013), pp. 226–239.

Lehmann, E.L. and Romano, J.P. *Testing Statistical Hypotheses*. 3rd ed. New York: Springer-Verlag, 2005.

Mehta, N., Rajiv, S. and Srinivasan, K. "Price Uncertainty and Consumer Search: A Structural Model of Consideration Set Formation." *Marketing Science*, Vol. 22(1) (2003), pp. 58–84.

Moraga-Gonzalez, J., Sandor, Z. and Wildenbeest, M. "Consumer Search and Prices in the Automobile Market." Unpublished manuscript, 2010.

Roberts, J. H. and Lattin, J. M. "Development and Testing of a Model of Consideration Set Composition." *Journal of Marketing Research*, Vol. 28(4) (1991), pp. 429–440.

Sorensen, A. (2000) "Equilibrium Price Dispersion in Retail Markets for Prescription Drugs." *Journal of Political Economy*, Vol. 108(4) (August 2000), pp. 833–850.

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