Testing Models of Consumer Search using Data on Web Browsing and Purchasing Behavior *

Babur De los Santos†
Ali Hortaçsu‡
Matthijs R. Wildenbeest§

First version: April 2009
This version: September 2009

Abstract

Using a large data set on web browsing and purchasing behavior we test to what extent consumers are searching in accordance to various classical search models. We find that the benchmark model of sequential search with a known distributions of prices can be rejected based on the recall patterns we observe in the data. Moreover, we show that even if consumers are initially unaware of the price distribution and have to learn the price distribution, observed search behavior for given consumers over time is more consistent with non-sequential search than sequential search with learning. Our findings suggest non-sequential search provides a more accurate description of observed consumer search behavior. We then utilize the non-sequential search model to estimate the price elasticities and markups of online book retailers.

Keywords: consumer search, electronic commerce, consumer behavior
JEL Classification: D43, D83, L13

*We thank Ken Hendricks for his useful comments and suggestions. In addition we thank seminar participants at the Industrial Organization Workshop of the NBER Summer Institute 2009 (Boston), the 2009 Far East and South Asia Meeting of the Econometric Society (Tokyo), and the SIEPR Conference on Internet Economics (Stanford). We gratefully acknowledge financial support from the NET Institute (www.netinst.org).

†Kelley School of Business, Indiana University, E-mail: babur@indiana.edu.
‡University of Chicago and NBER, E-mail: hortacsu@uchicago.edu.
§Kelley School of Business, Indiana University, E-mail: mwildenb@indiana.edu.

1 Introduction

Since Stigler's (1961) seminal paper, models of costly search have been at the heart of many economic models trying to explain imperfectly competitive behavior in product and labor markets. The theoretical literature typically models consumer search in two ways: following Stigler's original model, a strand of literature assumes *non-sequential* search behavior, where consumers sample a fixed number of stores, and choose to buy the lowest price alternative.[1] A much larger strand of the literature, starting with McCall (1970) and Mortensen (1970), points out that consumers cannot commit to a non-sequential search strategy in instances where the expected marginal benefit of an extra search exceeds the marginal cost. Thus, this literature argues that a sequential search model provides a better description of actual consumer search.[2]

Unfortunately, beyond the a priori reasons put forth by the literature, there have been few empirical studies of whether actual consumers follow sequential or non-sequential strategies. This is, no doubt, due to the difficulty of collecting data on individual search behavior. Therefore, most of what we know about individual level search behavior is from laboratory experiments. The majority of the experimental literature on search has focused on sequential search.[3] Schotter and Braunstein (1981) have reported that on average subjects tend to search in a fashion that is consistent with sequential search strategies, although subjects tend to search too little to be searching optimally. Kogut (1990) and Sonnemans (1998) find evidence that individuals are making decisions based on the total return from searching instead of on the marginal return from another draw as they would do if searching sequentially, resulting in too little search. Moreover, Kogut (1990) finds that in about a third of the time individuals accepted old offers, which violates optimal policy. Zwick et al. (2003) also find large rates of recall among participants of an experiment in which a randomly selected object with a known rank order has to be selected. Harrison and Morgan (1990) directly compare non-sequential and sequential strategies to so-called variable-sample-size strategies. The latter strategy is described in Morgan and Manning (1985) and is a generalization of both non-sequential and sequential search since it allows individuals to choose both sample size and how many times to search. Harrison and Morgan (1990) report that experimental subjects

[1] See also Burdett and Judd (1983) and Janssen and Moraga-González (2004).

[2] Examples of sequential search models in the consumer search literature are Axell (1977), Reinganum (1979), Carlson and McAfee (1983), Rob (1985), and Stahl (1989).

[3] See Camerer (1995, pp. 670-73) for a review of this literature.

indeed employ the least restrictive strategy if they are allowed to do so.

Aside from experimental studies, Hong and Shum (2006) and Chen, Hong, and Shum (2007) are the only papers that we are aware of that have attempted to discriminate between sequential and non-sequential search models using data from a real-world market. Hong and Shum (2006) collect data on textbook prices, and estimate structural parameters of search cost distributions (i.e. the demand parameters) that rationalize the prices set by competing firms. They find larger search-cost magnitudes for the parametrically estimated sequential search model than for the non-parametrically estimated non-sequential search model. Similar data is used in Chen, Hong, and Shum (2007) to conduct a nonparametric likelihood ratio test for choosing among the nonparametrically, moment-based non-sequential and parametrically estimated sequential search models. Although certain parameterizations of the sequential search model are found to be inferior, they conclude that it is difficult to distinguish between the non-sequential search model and the log-normal parameterization of the sequential search model in terms of fit.

This paper utilizes novel data on the web browsing and purchasing behavior of a large panel of consumers to test classical models of consumer search. Our data, described in some detail in Section 2, allows us to observe the online stores visited while shopping for a particular item, and which store the consumer decided to buy from. As pointed out by Kogut (1990) and as we will argue in more detail in Section 3 below, under the reservation price (utility) rule prescribed by the "benchmark" model of sequential search, a consumer always buys from the last store she visited, unless she has visited all stores in her choice set. In Section 4, using data on consumers shopping for books online, we find that this prediction is rejected by a large number of consumers in our data set.

In Section 3, we discuss the Rosenfield and Shapiro (1981) model, which relaxes the assumption that consumers "know" the distribution of prices while deciding on their search strategy, and allow for learning of the price distribution. Importantly, in this setting, the sequential search model can not be rejected based on recall patterns alone. Instead, we derive bounds on search costs that rationalize observed search behavior, and conduct tests based on the consistency of these search cost bounds across shopping trips. In Section 5, we explore whether misspecification of the search model is quantitatively important in our particular setting. In particular, we estimate consumer search cost distributions (the demand parameters) under various search rules. We find that the

estimated search costs under the non-sequential search assumption display much less dispersion within person than the search costs estimated under the sequential search with Bayesian learning model. This means the non-sequential search model leads to more stable parameter estimates, and we thus conclude that non-sequential search may provide a more accurate description of observed behavior.

Finally, in Section 7 we use the favored non-sequential search model to estimate the price elasticities faced by online retailers, and, under static profit maximization, the markups charged by these retailers. To do this, Section 6 derives expressions for demand elasticities implied by the non-sequential search model. One important feature of this model is that we allow for *asymmetric sampling*: due to for instance advertising or prior shopping experience, consumers' first draw may be skewed towards some online retailers (e.g. Amazon) over others.

Our results, reported in Section 7, indicate higher price elasticities than reported by Chevalier and Goolsbee (2003), especially for Amazon. A further discussion of our results vis a vis prior findings is in Section 7.1.

2 Data

We construct the dataset using two sources of data. The main data comes from the ComScore Web-Behavior Panel and includes detailed online browsing and transaction data from 100,000 Internet users for 2002 and 52,028 users for 2004. The users were chosen at random by ComScore from a universe of 1.5 million global users. ComScore is a leading provider of information on consumers' online behavior and supplies Fortune 500 companies and large news organizations with market research on e-commerce sales trends, website traffic, and online advertising campaigns. Each user's online activity is channeled through ComScore proxy servers that record all Internet traffic, including information on visits to a website or domain (browsing), as well as secure online transactions. The data include date, time, and duration of visit, as well as price, quantity, and description of each product purchased during the session.

We find that individuals in the ComScore sample are representative of online buyers in the United States. Comparing Internet users that have bought a product online on the sample with the Internet and Computer Use Supplement of the Current Population Survey (CPS) and the Forrester Technographics Survey, we find that the samples are similar in terms of the age, education, income,

household composition and other observable characteristics. The main differences of the ComScore sample, is that Internet users are older, with higher income, and more likely to be in college (those with "some college but no degree") than the CPS sample. The racial composition is similar across samples—online users are predominantly white. However, compared with CPS, ComScore oversamples Hispanics and Forrester oversamples whites. The geographic distribution of users is similar to CPS population estimates at the regional and state levels. Using the ComScore sample, we find that book buyers, those who purchased at least one book online, are slightly older, with greater income and more education than those who had any online transaction. We refer to De los Santos (2008) for a more detailed description of the sample.

The dataset contains users' transactions for products and services from June 2002 to December 2002 and for the full year of 2004. We excluded observations from firms that could not be identified as online bookstores, such as unidentified domains and auction sites. In total, 18 percent of the transactions were excluded; most of these were from Ebay.com (15 percent of transactions). Although the excluded transactions represent a large number of observations, they cannot be considered sales from an online bookstore because they are auctions of potentially different books, for example used books, autographed volumes, or auctioned items. A small number of transactions from international Amazon websites (in the United Kingdom, Canada, and Denmark) were also dropped. Given that Borders transaction were handled by Amazon in 2002 and 2004, we excluded browsing activity from Borders.com to avoid double counting.[4] Approximately 38 percent of the users realized a product transaction in 2002 (48 percent of users in 2004), and 7 percent of users bought at least one book online in 2002 (10 percent in 2004). This results in transactions from 15 online bookstores with 7,558 observations in 2002 and 8,020 observations in 2004.[5]

In order to analyze consumer search of online bookstores, we grouped small bookstores into two categories to create four firms: Amazon (66 percent of transactions), Barnes and Noble (20 percent), Book clubs (11 percent), and Other bookstores (4 percent). "Book clubs" include the following sites (.com): Christianbook, Doubledaybookclub, Eharlequin, Literaryguild, and Mysteryguild. Other bookstores include (.com): 1bookstreet, Allbooks4less, Alldirect, Booksamillion, Ecampus, Powells,

[4] Although initially Borders operated Borders.com, in April 2001 it signed a commercial agreement giving Amazon control of customer service, fulfillment, and inventory operations. As a result all visits to Borders.com are redirected to Amazon.com. In 2008 Borders relaunched Borders.com as an independent online bookstore.

[5] Each observation represents a single book purchased during one transaction; if multiple copies of the book are purchased in the same transaction, it is recorded as one observation.

Varsitybooks, and Walmart. Table 1 displays the number of transactions and visits per bookstore for the firm groups.

	Transactions		Visits	
Bookstore	Number	%	Number	%
Amazon	10,206	65.5%	249,593	76.3%
Barnes and Noble	3,046	19.6%	25,758	7.9%
Book Clubs				
christianbook.com	615	3.9%	3968	1.2%
doubledaybookclub.com	468	3.0%	4001	1.2%
eharlequin.com	61	0.4%	3647	1.1%
literaryguild.com	326	2.1%	3500	1.1%
mysteryguild.com	188	1.2%	2095	0.6%
Other Bookstore				
1bookstreet.com	10	0.1%	120	0.0%
allbooks4less.com	5	0.0%	199	0.1%
alldirect.com	27	0.2%	490	0.1%
ecampus.com	114	0.7%	1206	0.4%
powells.com	68	0.4%	1326	0.4%
varsitybooks.com	16	0.1%	218	0.1%
walmart.com	183	1.2%	28663	8.8%
booksamillion.com	245	1.6%	2290	0.7%
Total	15,578	100.0%	327,074	100.0%

Table 1: Transactions and Visits by Bookstore

The browsing activity of all users consists of 112,361 visits to the websites of online bookstores in 2002 and 214,713 visits in 2004.[6] In order to identify a user's visit to a website as search behavior related to a particular transaction, we link the browsing history up to 7 days before a transaction. There is no evidence to guide the definition of a search time span in relation to a transaction. One week is long enough to capture all search behavior related to a transaction; any longer intervals are likely to also capture unrelated website visits. A search history could be less than 7 days if another transaction has occurred within 7 days. Limiting browsing to search occurring 7 days prior to a purchase reduces the sample to 18,349 observations in 2002 and 25,513 in 2004. Although some user search may not be linked to the next transaction, but to a subsequent one, there is no clear way to link this intervening search to a later transaction. For example, if a user searches prices for book A but buys book B first, the search for book A is linked to book B. In the case where multiple books are acquired in the same transaction, browsing is linked to all books purchased. In the results we use several definitions of the relevant search period, from 7 days to the same day of

[6]This large increase was the result of a more than twofold increase in the number of visits to Amazon, which is the largest online bookstore and had 80 percent of website visits in 2004.

the transaction. Table 6 gives descriptive statistics of the sample.

	2002		2004	
	Mean	Std. Dev.	Mean	Std. Dev.
Duration of each website visit (in minutes)				
Visits not within 7 days of transaction	8.89	13.03	7.69	12.36
Visits within 7 days, excluding transactions	12.21	15.55	10.72	14.84
Visits within 7 days, excluding transactions	19.04	18.26	15.74	17.37
Transactions only	27.90	17.69	25.93	17.68
Total duration, excluding transaction visits	32.47	49.80	38.41	78.33
Total duration, including transaction visits	43.77	43.72	47.68	65.99
Number of firms searched	1.27	0.54	1.30	0.56
Number of books per transaction	2.38	2.10	2.20	1.95
Transaction expenditures (books only)	36.70	40.67	32.21	35.68
Number of books purchased	17,956		17,631	
Number of transaction sessions	7,559		8,002	
Number of visits within 7 days	18,349		25,513	
Number of visits not within 7 days	94,012		189,200	

Table 2: Descriptive Statistics of ComScore Book Sample

Despite the relative large number of online bookstores in 2002 and 2004, the market is highly concentrated, with the two dominant firms capturing 83 percent of the market: Amazon (66 percent of book sales) and Barnes and Noble (17 percent).[7] Amazon was visited in 74 percent book transactions, and in only 17 percent of transactions did Amazon buyers browse any other bookstore. Also, this firms capture most of the searching activity online. Of the 234 online bookstores listed on the Yahoo directory, the 15 bookstores in the sample captures 98.4 percent of all consumer visits to an online bookstores. The dominance of Amazon and Barnes and Noble in the market might explain the low levels of consumer search: users on average searched 1.2 bookstores in 2002 and 1.3 in 2004 (De los Santos, 2008).

Given the large number of online bookstores relative to the low number bookstores actually visited, we need to define which bookstores are relevant in the consumer search process as consumers might not be aware of all the online bookstores in the market. We construct consumers awareness of different bookstores by analyzing the consumer's browsing history within the dataset. For each transaction, a consumer is aware of a given bookstore if she has previously visited the bookstores. For a given search sequence the number of bookstores N is defined as the number of bookstores a consumer is aware at the time of the transaction. Figure 1 displays the distribution of consumer bookstore awareness.

[7] Books sales in dollars for 2004 from the ComScore data sample.

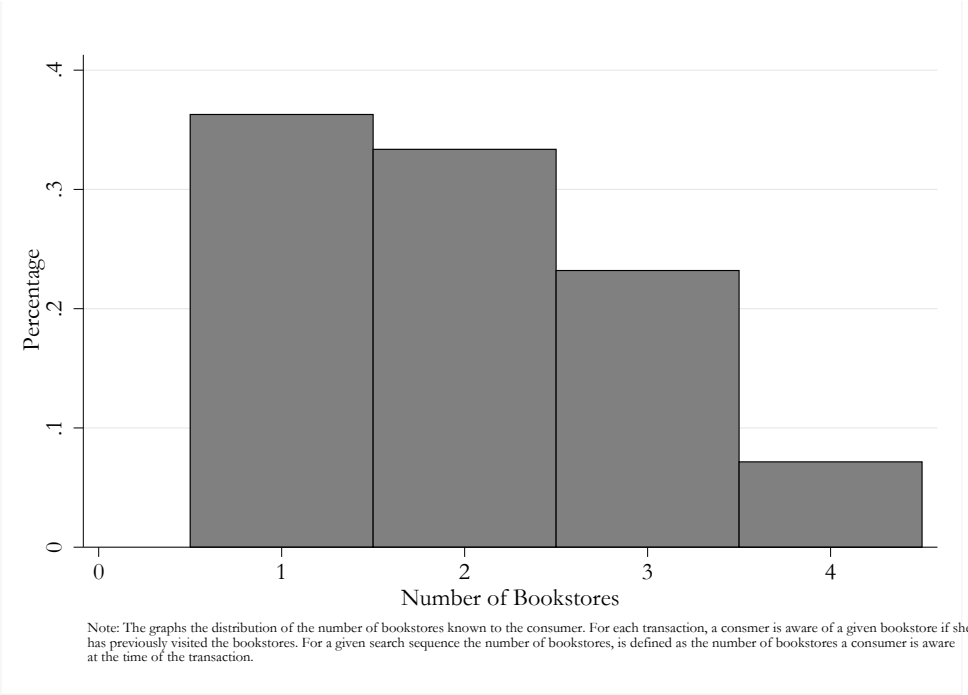


Figure 1: Consumer Bookstore Awareness

A limitation of the ComScore data is although we observe consumers visits to different retailer, we only observe the price of the transaction. We use two methods to recover missing prices for those visited bookstores. First, we use the most recent transaction prices at those bookstores with missing values. Second, we merged the book price information from a price comparison website to recover the distribution of prices.

This second data set contains more detailed information on prices and availability across stores for selected titles and is constructed using data from mySimon.com, a popular price comparison website.[8] By scouring thousands of web sites the search system of mySimon.com lists Internet stores in categories like computers, books, electronics, apparel, music, and movies. Each product listed on mySimon.com has a unique web page that gives the prices of online stores selling the item, as well as information on availability, store ratings, shipping costs, and sales taxes. We automatically collected this data for fourty-two book titles in the period between August and September 2004 using a web spider written in Java. The data set contains reference books, textbooks, as well (non-)fiction paperback and hardcover books. Although we only have information for a limited number of titles, a substantial share of the (non-)fiction book titles in our data set appeared on several of

[8]See http://www.mysimon.com.

The New York Times Best Sellers Lists during 2004.

In total fourteen different bookstores have posted prices on mySimon.com during the sampling period for at least on of the book titles.[9] Table 3 gives some summary statistics for the data.

	NYT bestseller fiction (hardcover)	NYT bestseller fiction (paperback)	NYT bestseller non-fiction	Random books	Reference books	Textbooks
Number of observations	2056	1542	1145	1818	335	554
Number of book titles	12	9	5	8	4	4
Number of stores	12.8	12.4	12.7	11.4	4.6	6.9
Unit price	13.94	7.73	13.50	10.67	225.31	68.69
Difference between *max* and *min*	6.20	5.38	5.52	6.20	34.18	15.12
Standard deviation	1.89	1.61	1.47	1.9	13.38	5.40
Coefficient of variation	13.76%	20.96%	10.73%	15.15%	5.2%	8.3%
Difference price BN and Amazon	0.69	0.25	1.27	2.04	-	-
% of obs. BN and Amazon equal	55%	54%	40%	28%	-	-
% of obs. BN more expensive	31%	41%	41%	64%	-	-
% of obs. Amazon more expensive	14%	6%	19%	7%	-	-

Notes: Barnesandnoble.com (BN) did not post any prices of the textbooks and reference books in our sample on mySimon.com during the sampling period.

Table 3: Summary Statistics

We use price data from both ComScore and mySimon.com to estimate the bounds of the price distribution, $\overline{p}, \underline{p}$. The prices from ComScore were the minimum and maximum transaction prices for a given product within the entire span of the dataset. MySimon tracks about 40 books during August and September of 2004 (8 books were not bought in the entire ComScore data), we use the minimum and maximum prices for this period. Since mySimon provides stocking information, we discard prices were the prices was not in stock, back-order, pre-order and other, or if it was flagged as refurbished.

3 Empirical Implications of Search Models

Consider first the classic sequential search model of McCall (1970), in where a consumer is sampling stores selling a homogenous good. Assuming the consumer believes that each store's price is an i.i.d. draw from distribution $F(p)$ (which may arise as the result of a symmetric mixed strategy pricing equilibrium), the consumer will continue to search as long as she finds a price greater than

[9] We exclude Amazon Marketplace, which is reported by mySimon in some instances, since the prices were likely be from used books, for 80 percent of the books marketplace was the lowest price of the bookstores for prices such as 1 cents to a dollar.

some reservation price $r(c)$, where $r(c)$ is given by:

$$c = \int_{\underline{p}}^{r(c)} (r(c) - p) f(p)\, dp. \tag{1}$$

As seen in the equation, the reservation price is such that, if the price in hand is $r(c)$, the marginal cost of search c equals the expected benefit from continuing searching. (The integral on the right-hand side is the expected reduction in price from another search, accounting for the option value of discarding higher price draws.)

Since for a consumer with search cost c the reservation price $r(c)$ is constant across searches, the consumer will never recall a price that she sampled earlier, unless there are a finite number of stores, and the consumer has visited all the stores. Our first test of the sequential search hypothesis will thus focus on recall behavior by consumers.

Test 1 (No Recall) *Under the null hypothesis of the McCall model, we should not observe recall of already sampled alternatives, unless the consumers has exhausted sampling all of the stores whose existence she is aware of.*

Note that the above test would apply if consumers were considering attributes other than price, i.e. we could have rewritten the model in which firms were offering a distribution $F(u)$ of net utilities, and the stopping rule would have been couched in terms of a reservation utility level. Moreover, the test allows for unequal sampling probabilities across firms.

Observe that the absence of recall in the model described above depends crucially on the constant reservation price rule. We now discuss a variant of the sequential search model that may lead to a sequence of reservation prices that are increasing in the number of prices sampled. This is the model of Rosenfield and Shapiro (1981), in which consumers learn about the price distribution while sampling. More specifically, we assume searchers learn by Bayesian updating their Dirichlet priors over prices that follow a multinomial distribution.[10]

In particular, suppose a random vector of prices $p = (p_1, \ldots, p_n)$ follows a multinomial distribution with probabilities of each price $\pi = (\pi_1, \ldots, \pi_k)$. The vector π containing the true probabilities of each price being charged is unknown, but its prior distribution is assumed to be Dirichlet with

[10]Since the Dirichlet distribution is the conjugate prior of the multinomial distribution the posterior distribution will be Dirichlet as well. This means this combination of distribution and prior is relatively easy to work with and allows us to retrieve simple expressions.

parameters $\alpha = (N_1, \ldots, N_n)$, where α_i can be interpreted as the frequency of price p_i, i.e., N_i/N, where $N = \sum N_i$. Then the posterior distribution of π after observing a price p_i is a Dirichlet distribution with parameters $\alpha^* = \alpha + e_i$, where e_i is a vector that contains 1 on the ith place and 0 everywhere else.[11] For instance, suppose consumers have an uninformative uniform Dirichlet prior distribution, i.e., $\alpha = (1, 1, 1, \ldots, 1)$. If a consumer observes a price p_2 we simply add 1 to α_2 to get a Dirichlet posterior distribution with parameters $\alpha^* = (1, 2, 1, \ldots, 1)$. Note that the impact of an additional price observations on the posterior is decreasing in N, which means we can interpret N as an indicator of how certain searchers are about their prior.

As shown by Rosenfield and Shapiro (1981) in the above setting consumers' optimal search policy is myopic and can be characterized by a reservation price that is non-decreasing in the number of prices sampled. This means that unlike the standard sequential search model a sequential search model with Bayesian updating can in fact explain why some consumer return to previously visited stores, even if they have not exhausted all their search possibilities.

An attractive feature of the model is that the gains from searching only depend on how sure consumers are about their prior and on the lowest observed price so far. Moreover, if a consumer returns to a previously visited store to buy the good, that is, the consumer recalls, the range of search costs that rationalizes this behavior is relatively small. To see this, suppose there are N possible prices $\mathbf{p}$: $p_1 \leq p_2 \leq \cdots \leq p_N$. For simplicity, assume the consumer's prior after having observed some initial price p_k is uninformative, i.e., $N_i = 1$. The gains from search $G(\cdot)$ at the lowest observed price so far, denoted p_N^*, are then

$$\begin{aligned} G(p_N^* = p_k|\mathbf{p}) &= p_k - \left(\sum_{i=1}^{k-1} \frac{1}{N} \cdot p_i + \frac{N-(k-1)}{N} \cdot p_k \right) \\ &= \frac{k}{N} \cdot p_k - \sum_{i=1}^{k} \frac{1}{N} \cdot p_i. \end{aligned}$$

Intuitively, the gain is equal to the price at hand minus the expected price when searching, using that a consumer will stick to the current price in the unfortunate event a higher price than p_k is sampled. After having observed a second price the searcher will update her prior. We are interested in recall patterns, so suppose this second price is higher than p_N^*. Since the gains of search only depend on the probability of finding a price lower than p_N^*, when the second price is higher than

[11] See Theorem 1 on p.174 of DeGroot (1970).

p_N^* the gains from search will only be affected through the change in N:

$$\begin{aligned} G(p_{N+1}^* = p_k|\mathbf{p}) &= p_k - \left(\sum_{i=1}^{k-1} \frac{1}{N+1} \cdot p_i + \frac{N+1-(k-1)}{N+1} \cdot p_k\right) \\ &= \frac{k}{N+1} \cdot p_k - \sum_{i=1}^{k} \frac{1}{N+1} \cdot p_i \\ &= \frac{N}{N+1} G(p_N^* = p_k|\mathbf{p}). \end{aligned}$$

Therefore, if we observe a consumer searching once more, but returning to the store visited before the additional search to buy the good, we know her search cost should have been:

$$\frac{N}{N+1} G(p_N^*|\mathbf{p}) < c < G(p_N^*|\mathbf{p}). \tag{2}$$

To illustrate what this implies for the setting we are studying, suppose we observe a consumer searching online for a particular book. The consumer is first going to Amazon, then to Barnes and Noble, but finally buys the book at Amazon for a price of \$7. To be able to calculate the search cost that rationalizes the observed behavior, given the above search protocol, we need to make an assumption about the support of the prior as well as how much weight the searcher puts on her prior, which is captured by N. For the latter we take the number of online stores selling books the consumer is aware of, while we use the observed support of the empirical price distribution for the book to approximate the support of the consumer's prior. Suppose this support is $[\underline{p}, \overline{p}] = [6, 12]$, the consumer has observed an initial price $p_N^* = 7$ at Amazon, and suppose the consumer is aware of 5 bookstores, so we set $N = 5$. Assuming the consumer's prior distribution after having visited Amazon is uniform the gains from searching are

$$\begin{aligned} G(p_N^* = 7) &= p_N^* - \left(\frac{p_N^* - \underline{p}}{\overline{p} - \underline{p}} \cdot \frac{p_N^* + \underline{p}}{2} + \frac{\overline{p} - p_N^*}{\overline{p} - \underline{p}} \cdot p_N^*\right); \\ &= 7 - \left(\frac{1}{6} \cdot 6\frac{1}{2} + \frac{5}{6} \cdot 7\right) = 0.08, \end{aligned}$$

which is just the price at hand minus the expected price when searching, assuming a continuous uniform distribution. Using equation (2) and the weight put on the prior, we know the search cost c of this consumer should have been $c \in (0.07, 0.08)$. Note we only need the bounds of the price distribution and the transaction price for this calculation, and not the price at Barnes and Noble.

As illustrated in the example above, to rationalize recall behavior for searchers searching sequentially with Bayesian updating requires very specific search cost values. This means if we observe

consumers recalling, given the support of the price distribution and weight put on the prior, we can be pretty specific in what their search cost should have been. As discussed in more detail in Section 5, this will be used to test whether sequential search with learning can explain the recall patterns we observe.

By definition recalling does not put any restrictions on the search costs of consumers who are searching non-sequentially. To see this, define c_k as the search cost of a consumer who is indifferent between searching k and $k+1$ times, i.e.,

$$c_k = E[\min_k p] - E[\min_{k+1} p].$$

A consumer who finds it optimal to sample k firms should then search cost c such that $c_k < c < c_{k-1}$, no matter whether the consumer recalls ore not. If we stick to that assumptions that consumers have an uninformative prior with upper bound $\overline{p}$ and lower bound $\underline{p}$, this equation simplifies to

$$c_k = \frac{\overline{p} - \underline{p}}{k^2 + 3k + 2}.$$

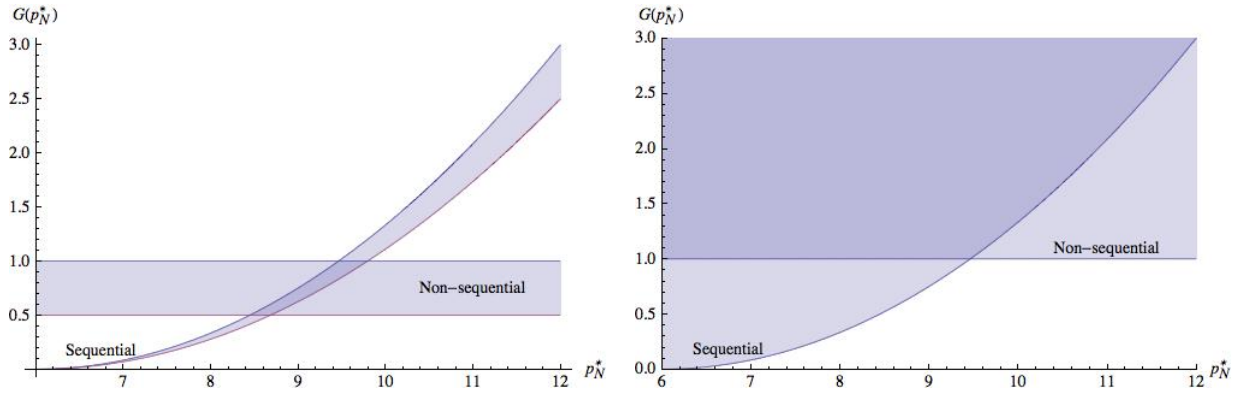
This means a consumer who finds it optimal to search twice should have search cost such that $c_2 < c < c_1$, i.e.,

$$\frac{\overline{p} - \underline{p}}{12} < c < \frac{\overline{p} - \underline{p}}{6}.$$

Note that because of the uniform prior we only need the upper and lower bound of the empirical price distribution to derive the bounds on search costs that rationalize observed search patterns. For instance, in the online bookstore example above, if the consumer were sampling two stores non-sequentially instead of sequentially search cost should have been $c \in (0.50, 1.00)$.

Figure 2(a) shows how the estimated search cost intervals for sequential search with Bayesian updating and non-sequential search relate to each other for the online bookstore example. The graph shows the gains from search as a function of the lowest observed price after having searched once and twice. Notice that the number of searches is determined before the actual search starts in the non-sequential search model, so there is no learning going on: the search cost range that rationalizes behavior is therefore independent of the transaction price. In the sequential search model with Bayesian updating the search cost and search cost range are increasing in the transaction price.

Now let's look how other search patterns can be rationalized by the sequential search model with Bayesian updating. If a consumer searches just once search cost should be higher than the



(a) Searching twice and buying from first firm

(b) Searching once

Figure 2: Gains from search as a function of the transaction (and lowest observed) price p_N^*

gains from continued search, so

$$c > G(p_N^*|\mathbf{p}).$$

In the non-sequential search model search cost should be higher than the critical search cost value c_1, where c_1 is again the search cost of a consumer who is indifferent between searching once and twice. Figure 2(b) shows what range of search cost values rationalize searching once in both settings for the online bookstore example. In both settings there is no upper bound on the search cost that rationalizes searching once, which means that the two regions overlap for a large set of search values.

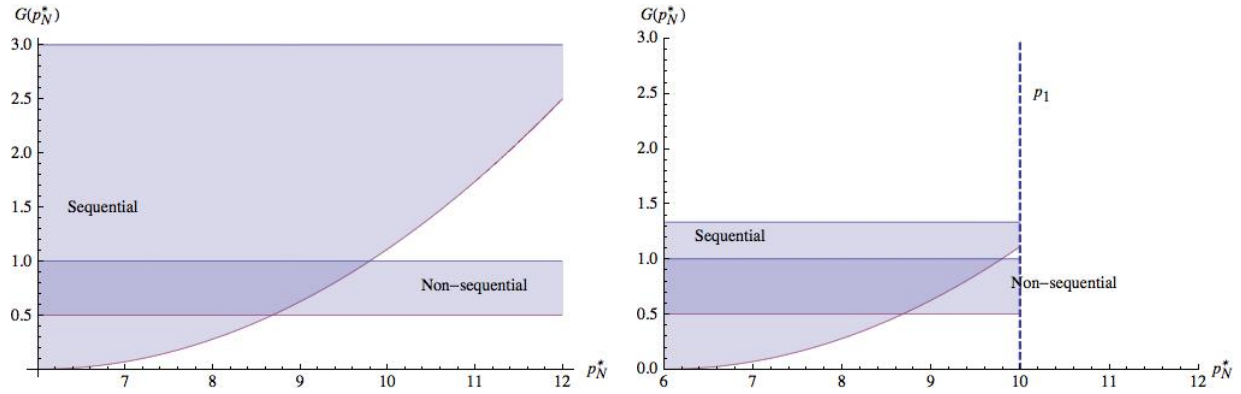
Now suppose a consumer searches twice and buys from the second firm. While the non-sequential search case is not different from a situation where the consumer buys from the first firm, under sequential search with Bayesian updating the regions will change. If we only observe the transaction price all we can say is

$$\frac{N}{N+1}G(p_N^*|\mathbf{p}) < c < G(\overline{p}|\mathbf{p}).$$

Figure 3(a) again shows the region of search cost that can rationalize behavior as a function of the transaction price p_N^* for both the sequential and non-sequential case. However, if we also observe the first price p_1, then

$$\frac{N}{N+1}G(p_N^*|\mathbf{p}) < c < G(p_1|\mathbf{p}).$$

Figure 3(b) plots the regions when the first price observed is $p_1 = 10$. Note that by definition the transaction price p_N^* is lower than p_1.



(a) Searching twice and buying from second firm (only transaction price p_N^* observed)

(b) Searching twice and buying from second firm (both prices observed, first price $p_1 = 10$)

Figure 3: Gains from search as a function of the transaction (and lowest observed) price p_N^*

To summarize, for consumers who search twice but buy from the first firm we can obtain pretty precise estimates of their search costs, assuming they search sequentially with Bayesian updating. In most other cases the region of search costs that rationalize observed behavior are going to be much larger.

4 Testing the "no recall" hypothesis

The benchmark sequential search model of McCall tells us that the only instance that a consumer will recall a store is if she exhausts the search by visiting all firms. If the consumer does not exhaust the search, the optimal stopping rule is to buy from the last firm visited (i.e. price is below reservation price).

To test this hypothesis, we have to check whether (i) a consumer recalled a product that was previously sample, and (ii) if there was a recall, whether this was because the consumer exhausted her search over all retailers she is aware of. To do this, we first identify all the stores that a consumer is aware of by looking at previous visits to bookstores by that consumer. E.g., if we observe that the consumer has only visited Amazon and BN in the past, this is a conservative lower bound on the set of retailers that the consumer is aware of.

For a given transaction the consumer visits one store or the consumer searches more than one store. If the consumer visits more than one store, she either buys from the last store, or she recalls a previously visited store. In the case where the consumer visits one firm, we cannot distinguish between sequential and non-sequential strategies.

Table 4 shows the percentage of transactions for each of the three search sequences for different definitions of the search period considered. The periods range from one week prior to the same day of the transaction. For example, for the search period defined as the same day of the transaction (bottom row of the table), in 90 percent of the transactions the consumer visited one firm in the same day. In 10 percent of transactions, the consumers visited more than one bookstore. Among the 10 percent of transactions in which a consumer visited more than one store, 62 percent bought from the last firm sampled and 38 percent recalled a previously visited firm.

Search window	No. of visited visited		If 2 or more firms, bought from:		Exhausted search?
7 Days	One	76%			
	2 or more	24%	Last firm sampled	65%	
			Recalled	35%	55%
6 Days	One	77%			
	2 or more	23%	Last firm sampled	64%	
			Recalled	36%	55%
5 Days	One	79%			
	2 or more	21%	Last firm sampled	63%	
			Recalled	37%	55%
4 Days	One	80%			
	2 or more	20%	Last firm sampled	61%	
			Recalled	39%	55%
3 Days	One	82%			
	2 or more	18%	Last firm sampled	61%	
			Recalled	39%	56%
2 Days	One	84%			
	2 or more	16%	Last firm sampled	61%	
			Recalled	39%	56%
1 Day	One	86%			
	2 or more	14%	Last firm sampled	61%	
			Recalled	39%	56%
Same day	One	90%			
	2 or more	10%	Last firm sampled	62%	
			Recalled	38%	58%

Table 4: Test of "no recall" hypothesis

Note that there are a large number of instances where the consumer recalls a product that was previously sampled. This may not immediately be construed as evidence against a sequential model, however, as recall is allowed in a sequential search in which a consumer has exhausted the search options available to her. The last column presents the percentange of the transactions where the search were exhausted for each search sequence. Exhausting the search means that the consumer searched all the firms they know (have visited before) at the time of the transaction. If we focus on the bottom row of the table, where we look at search activity only on the day of the transaction, we see that consumers "exhausted" the search possibilities in 58 percent of those transactions where they recalled a previously sampled product. Perhaps, more to the point, consumers did not exhaust

the search in 42% of the recalled instances, which is a violation of the basic sequential search model. Note that our definition of "not exhausting a search" is a conservative one; it may have been the case that the consumer was aware of more bookstores than we were able to capture with our data set.

Search Window	Total	Amazon	Barnes & Noble	Book Clubs	Other bookstores
Recall percentage by firm					
7 Days	35%	50%	20%	18%	23%
6 Days	36%	51%	22%	20%	24%
5 Days	37%	52%	23%	23%	26%
4 Days	39%	54%	24%	27%	27%
3 Days	39%	54%	25%	28%	27%
2 Days	40%	54%	26%	28%	28%
1 Day	40%	54%	26%	31%	27%
Same Day	39%	53%	25%	30%	26%
Distribution of Recall Transactions					
7 Days	1302	70%	18%	6%	6%
6 Days	1283	68%	20%	7%	6%
5 Days	1230	67%	20%	7%	6%
4 Days	1187	66%	20%	8%	6%
3 Days	1103	65%	21%	8%	7%
2 Days	994	64%	22%	7%	7%
1 Day	845	63%	22%	7%	7%
Same Day	588	64%	22%	6%	8%

Table 5: Recall by Firm

Table 5 looks in more detail at the recall transactions by linking recalls to the bookstores where the final transaction took place. The table shows that in most cases searchers recalled to Amazon and Barnes and Noble: only in 14% of the transactions in which consumers recalled a previously visited firm on the same day of the transaction they recalled a book club or a bookstore from the other bookstores category.[12] Moreover, Table 5 also shows that Amazon.com visitors are much more likely to recall than visitors of other bookstores: on the transaction day 53% of Amazon buyers have recalled, while this is only between 25 and 30% for the other bookstores.

5 Estimates of search costs implied by sequential and non-sequential search models

The results of the previous section rule out the basic sequential search model with a constant reservation price strategy. However, as we argued above, once we allow for Bayesian updating, the observation recall no longer invalidates sequential search. Thus we proceed with an alternative

[12]Note that some of the recall transactions in the book clubs and other bookstores categories might be to a different bookstores within the same group. As Table 5 shows, given the small percentages this will not have a major impact on our results.

testing strategy in this section: we will estimate search cost bounds implied by both non-sequential and the sequential with Bayesian updating models. Since this will yield multiple search cost bounds for a given consumer in our data set, we will then check whether one model yields more consistent search cost bounds across transactions for a given customer.

5.1 Bounds generated by the Rosenfield-Shapiro model

Recall from Section 3, that under the Rosenfield-Shapiro model, if we observe a consumer searching twice, but buying from the first firm we know his search cost c should be bounded between:

$$\frac{N}{N+1}G(p_N^*|\mathbf{p}) < c < G(p_N^*|\mathbf{p})$$

where, assuming a continuous uniform prior, the gains from search after having observed an initial price p_N^* are

$$G(p_N^* = p) = p_N^* - \left(\frac{p_N^* - \underline{p}}{\overline{p} - \underline{p}} \cdot \frac{p_N^* + \underline{p}}{2} + \frac{\overline{p} - p_N^*}{\overline{p} - \underline{p}} \cdot p_N^*\right) = \frac{p_N^* - \underline{p}}{\overline{p} - \underline{p}} \cdot \frac{p_N^* - \underline{p}}{2}.$$

To estimate the model we need N, the bounds of the price distribution $\overline{p}, \underline{p}$ and the transaction price. N is the number of firms known to each consumer at the time of the transaction. As before, our empirical definition of when a consumer "knows" a store is if she has visited it prior to the transaction within the span of the dataset.

To estimate the bounds of the price distribution, $\overline{p}, \underline{p}$, we use price data from comScore and mySimon.com. The prices from comScore were the minimum and maximum transaction prices for a given product within the entire span of the dataset. mySimon tracks 42 books during August and September of 2004 (8 books were not bought in the entire 2002,2004 comScore data), we use the minimum and maximum prices for this period. Since my simon provides stocking information, we discard prices were the prices was not in stock, backorder, pre-order and other, or if it was flagged as refurbished. We excluded Amazon Marketplace, which is reported by mySimon, since the prices were likely to be for used books. For 80 percent of the books Amazon Marketplace was the lowest price among all bookstores with prices as low as 1 cent to a dollar.

For people that searched and bought from one firm, we only observe the lower bound:

$$c_l^s = G(p_N^*|\mathbf{p}).$$

Unfortunately, we can not calculate an upper bound for consumers who visited only one store.

For searchers who visited more than one firm (recalled or bought from the last firm), the lower bound of the search cost is

$$c_l^s = \frac{N}{N+1} G(p_N^* | \mathbf{p}).$$

The upper bound of the search cost for those who searched more than one firm and recalled is

$$c_u^s = G(p_N^* | \mathbf{p}).$$

For people who searched more than one firm and bought from the last one, the upper bound on the implied search cost is:

$$c_u^s = G(\overline{p} | \mathbf{p}) = G(p_N^* = p) = \frac{\overline{p} - \underline{p}}{2}.$$

5.2 Bounds generated by the non-sequential search model with uniform distribution

For a non-sequential model with an uniform distribution, the cutoffs of the search cost are given by

$$c_k^{ns} = \frac{\overline{p} - \underline{p}}{k^2 + 3k + 2}.$$

where k is the number of firms searched. For a transaction the bounds are

$$c_l^{ns} = \frac{\overline{p} - \underline{p}}{k^2 + 3k + 2} \leq c \leq \frac{\overline{p} - \underline{p}}{(k-1)^2 + 3(k-1) + 2} = c_u^{ns}.$$

We do not observe the upper bound for those who only visit one firm, $k = 1$.

5.3 Results

For each search session that ended in a purchase, we estimate the lower and, whenever possible, upper bounds for both search strategies. Given an upper and lower bound on the search cost implied by the data, we calculate the midpoint of the bounds as our point estimate of the search cost.[13]

We then calculate the within consumer standard deviation of the "midpoint" search cost estimates. Figure 4 displays the within-consumer standard deviation of search costs implied by the sequential vs. non-sequential models (although the "midpoint" is a somewhat arbitrary summary of the bounds, the figure does not change qualitatively if we plot the standard deviations of the

[13]Since neither model allows us to calculate an upper bound on the search cost when the consumer samples and purchases from a single store, we omit these observations from our calculations.

bounds separately). For both sequential and non-sequential search the sample consists of consumers who recall a previously visited store, while not having visited all stores they are aware of.

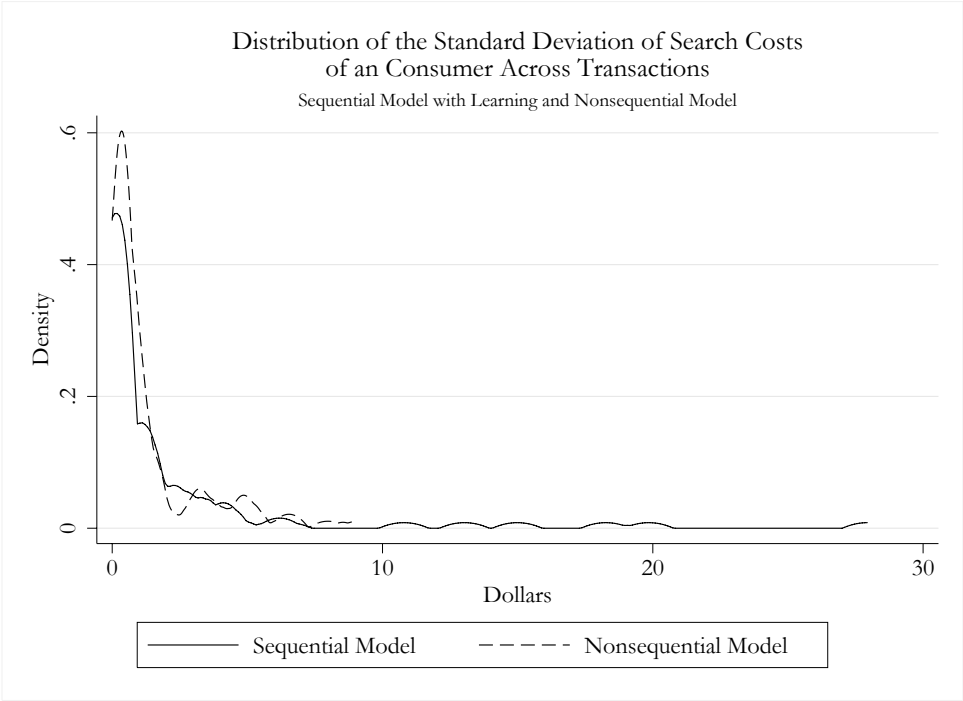


Figure 4: Dispersion of search costs

Observe that our estimates of search costs based on the non-sequential model display smaller within-person dispersion than our estimates based on the sequential model, although the differences are small. If we believe search costs to be relatively time invariant, the figure suggests that the non-sequential model does a better job explaining our data with a parsimonious model of behavior.

6 Implications of the nonsequential search model

We will now investigate price elasticities and (static) profit-maximizing firm behavior in an environment where consumers search non-sequentially. Based on the patterns we observed in our data, we allow for unequal *first* sampling probabilities and marginal cost heterogeneity.

There are N firms selling a good j at a price p_j. We will simplify matters by assuming consumers observe the empirical cumulative distribution function of stores' prices. This means consumers know

which prices are around, but do not know which store is offering what prices. Furthermore, we assume consumers obtain the first observation for free.

We assume consumers search non-sequentially and perfect recall, so consumers determine before they start searching how many times to search. The first sampling probability is denoted ρ_j and can be different across firms. For example, if Amazon has $\rho = 0.65$ this means 65% of consumers start their search there. For simplicity, we assume all subsequent sampling probabilities are similar across firms, i.e., conditional on searching twice, a consumer who has started searching at Amazon is equally likely to go to Barnes and Noble as to 1bookstreet.

The assumption that consumers observe the empirical price distribution function allows us to label the N stores by descending prices, $p_1 > \cdots > p_N$, which means the lowest ranked firm (store 1) offers the worst deal in terms of prices, while the highest ranked firm (store N) is offering the best deal. We can use this ordering to define α_{jk} as the probability the j-lowest ranked firm offers the lowest price out of k draws. To calculate α_{jk}, we consider two different sampling protocols. The first is sampling with replacement, which is in line with most of the search literature and the models in the previous section. The second is sampling without replacement, which we believe is more realistic given the setting.

Consider first the sampling with replacement case. Start with just one draw, i.e., $k = 1$. In this case all what matters is the probability of being sampled first, which means $\alpha_{j1} = \rho_j$. If $k = 2$ there will be two firms in the sample, which means the store offering the highest overall price will only offer the lowest price in the sample if it sampled twice, i.e., $\alpha_{12} = \rho_1/N$, where $1/N$ is the sampling probability beyond the first search. The second-lowest ranked store will only offer the lowest price when either this store is sampled twice or when it is sampled together with the lowest-ranked firm, which means $\alpha_{22} = (\rho_1 + 2\rho_2)/N$. Similarly, the probability the j-lowest ranked store will offer the lowest price in the sample of two is $\alpha_{j2} = (\rho_1 + \ldots + \rho_{j-1} + j\rho_j)/N$. More generally, as shown in the Appendix, we can use combinatorics to derive the probability that the j-lowest ranked firm offers the lowest price out of $k = 3$ random draws or more, i.e.,

$$\alpha_{jk} = \rho_j \left(\frac{j}{N}\right)^{k-1} + (\rho_1 + \ldots + \rho_{j-1}) \left(\left(\frac{j}{N}\right)^{k-1} - \left(\frac{j-1}{N}\right)^{k-1} \right). \tag{3}$$

Consider now the sampling without replacement case. A crucial difference with sampling with replacement is that stores can only be sampled once. This means the j-lowest ranked firm will never

be the one offering the lowest price in samples of size $k > j$. As before, with just one draw $\alpha_{j1} = \rho_j$, but now the lowest ranked store will never offer the lowest price in a sample of two, i.e., $\alpha_{12} = 0$. The second-lowest ranked store will only offer the lowest price when either this store is sampled first and the lowest ranked store second, which happens with probability $\rho_2/(N-1)$, or the other way around, which happens with probability $\rho_1/(N-1)$. This means $\alpha_{22} = (\rho_1 + \rho_2)/(N-1)$. Similarly, the probability the j-lowest ranked store will offer the lowest price in the sample of two is $\alpha_{j2} = (\rho_1 + \ldots + \rho_{j-1} + (j-1)\rho_j)/(N-1)$. More generally, as shown in the Appendix, we can use combinatorics to derive the probability that the j-lowest ranked firm offers the lowest price out of $k = 3$ random draws or more, i.e.,

$$\alpha_{jk} = \begin{cases} \left((\rho_1 + \ldots + \rho_{j-1})\frac{k-1}{j-1} + \rho_j\right)\frac{(j-1)\times\ldots\times(j-(k-1))}{(N-1)\times\ldots\times(N-(k-1))} & \text{if } j \geq k; \\ 0 & \text{if } j < k. \end{cases} \tag{4}$$

Equations (3) or (4) can be used to characterize optimal consumer behavior as well the supply side of the market. Consider first the consumer side of the market. Consumers are characterized by a search cost value which is drawn from a search cost distribution $G(c)$ with density function $g(c)$. The non-sequential search assumption allows us to define the critical search cost value c_k as the search cost of a consumer who is indifferent between searching k and $k+1$ times, i.e.,

$$c_k = E[\min_k p] - E[\min_{k+1} p].$$

Using probabilities α_{jk} the expected minimum price when searching k times is

$$E[\min_k p] = \sum_{j=1}^{N} \alpha_{jk} p_j,$$

which means we can write the search cost cutoffs as

$$c_k = \sum_{j=1}^{N} (\alpha_{jk} - \alpha_{j(k+1)}) p_j. \tag{5}$$

Consumers with search costs between c_{k-1} and c_k will search k times, so we can define μ_k as the share of consumers searching k times, i.e.,

$$\mu_1 = 1 - G(c_1) \text{ for } k = 1; \tag{6a}$$

$$\mu_k = G(c_{k-1}) - G(c_k) \text{ for } k = 2, 3, \ldots, N, \tag{6b}$$

where $G(c_N) = 0$ by assumption.

Next consider the supply side of the market. We can use the probabilities α_{jk} defined in equations (3) or (4) and the grouping of consumers μ_k given in equations (6a) and (6b) to calculate the market shares, i.e., the market share equation for store j is just the sum of the probability of selling to the different groups of consumers, multiplied by their shares in the consumer population:

$$q_j = \sum_{k=1}^{N} \alpha_{jk}\mu_k. \tag{7}$$

Store j's profits are given by

$$\Pi_j = Sq_j(p_j - mc_j),$$

where S is the size of the market and mc_j is firm j's marginal cost. Firms' static profit maximizing behavior implies the first-order condition for p_j should hold, i.e.,

$$q_j + (p_j - mc_j)\frac{\partial q_j}{\partial p_j} = 0. \tag{8}$$

In the Appendix we show the derivatives of the market share equations (7) are

$$\frac{\partial q_j}{\partial p_j} = -\sum_{k=1}^{N-1} (\alpha_{jk} - \alpha_{j(k+1)})^2 g(c_k). \tag{9}$$

6.1 Estimation

We observe sampling probabilities and prices, so we can directly calculate the c_k's defined in equation (5). We also observe μ_k, the shares of consumers searching k times, from which we can calculate $G(c_k)$ for $k = 1, 2, \ldots, N$ using equations (6a) and (6b). Combining the two gives a non-parametric estimate of the search cost cumulative distribution function (see also De los Santos, 2008).

From observed sampling probabilities ρ_j and the share of consumers searching k times μ_k we can get an estimate of the market shares by using equation (7). Equation (9) can be used to estimate the derivates of the market shares. However, the search cost PDF evaluated at the cutoffs are not observed, so we proceed by using the trapezoid method (see also Hortaçsu and Syverson, 2004) to derive an approximation, i.e.,

$$g(c_{k-1}) + g(c_k) = \frac{2[G(c_{k-1}) - G(c_k)]}{c_{k-1} - c_k} = \frac{2\mu_k}{c_{k-1} - c_k}.$$

Notice that in this case $g(c_0)$ is not identified, so we set it equal to zero. The estimates of $g(c_k)$ allow us to calculate marginal cost as

$$mc_j = p_j + \frac{q_j}{\partial q_j / \partial p_j}. \tag{10}$$

7 Application: price elasticities and markups of online bookstores

In this section, we present estimates price elasticities and markups of the online bookstores that appear in our sample using the model developed in the previous section. To estimate the model, we use our data set on search behavior completed with prices from the mySimon.com price database.[14] For four books that appear in the mySimon.com price database we have a sufficient number of transactions, so we focus on these books only. These books all have appeared on the New York Times Bestseller list for at least some period in 2004.

		Prices ($)				Consumers by Sample Size (%)		
Product name	Obs.	Mean	Std. Dev.	Min	Max	μ_1	μ_2	μ_3
The Da Vinci Code	48	14.48	0.49	13.91	14.97	0.751	0.227	0.022
The Five People you Meet in Heaven	24	11.70	0.28	11.34	11.97	0.756	0.244	0.000
The Rule of Four	17	14.60	1.55	11.97	15.88	0.846	0.154	0.000
R is for Ricochet	23	16.64	2.12	13.07	18.45	0.802	0.161	0.036

Notes:

Table 6: Descriptive Statistics

Table 6 gives descriptive statistics for the four books. Mean prices are similar across books, with *The Five People you Meet in Heaven* being a bit lower priced on average than the other books, while *R is for Ricochet* is priced a bit higher. The reported shares of consumers sampling k stores shows little variation across the books. In line with findings for the complete sample, consumers search activity is very modest: between 75% and 85% of consumers visits at most one bookstore before buying and only for two of the books consumers search more than twice.

Table 7 gives the estimated cutoff values of the search cost distributions that rationalize observed search patterns for the case of homogenous goods. These cutoff search costs are estimated using equation (5). We report our findings for sampling with replacement as well sampling without replacement. We allow for asymmetric first sampling probabilities – the probability a bookstore

[14] Book clubs did not appear on mySimon.com during the sampling period, so for this category we use weighted median transaction prices.

Product name	Cutoff search costs			CDF values			PDF values		
	c_1	c_2	c_3	$G(c_1)$	$G(c_2)$	$G(c_3)$	$g(c_1)$	$g(c_2)$	$g(c_3)$
Sampling with replacement									
The Da Vinci Code	0.404	0.213	0.115	0.249	0.022	0.000	1.939	0.440	0.000
The Five People you Meet in Heaven	0.228	0.125	0.071	0.244	0.000	0.000	4.734	0.000	0.000
The Rule of Four	0.931	0.629	0.442	0.154	0.000	0.000	1.021	0.000	0.000
R is for Ricochet	1.355	0.888	0.608	0.198	0.036	0.000	0.431	0.259	0.000
Sampling without replacement									
The Da Vinci Code	0.539	0.300	0.060	0.249	0.022	0.000	1.716	0.181	0.000
The Five People you Meet in Heaven	0.304	0.182	0.059	0.244	0.000	0.000	3.979	0.000	0.000
The Rule of Four	1.241	1.056	0.930	0.154	0.000	0.000	1.662	0.000	0.000
R is for Ricochet	1.807	1.464	1.194	0.198	0.036	0.000	0.672	0.268	0.000

Notes:

Table 7: Empirical Non-Sequential Search Cost CDF

is sampled first is estimated using all transactions in the database.[15] Also reported are the corresponding quantiles of the search cost distribution which are calculated using $G(c_k) = 1 - \sum_{i=1}^{k} \mu_i$. As explained in the previous section, given estimates of cutoff search costs c_k and corresponding CDF quantiles $G(c_k)$ we can calculate the values of the search cost PDF evaluated at the cutoff search costs using the trapezoid method. Estimated PDF values are displayed in the last three columns of Table 7.

Product name	Marginal Costs mc ($)				Elasticities E			
	Amazon	B&N	Book clubs	Other	Amazon	B&N	Book clubs	Other
Sampling with replacement								
The Da Vinci Code	11.72	10.48	12.50	12.50	-4.61	-3.34	-9.90	-8.75
The Five People you Meet in Heaven	10.52	10.09	10.65	10.86	-8.24	-6.36	-13.12	-23.43
The Rule of Four	8.47	6.44	9.44	12.47	-2.20	-1.68	-4.73	-5.82
R is for Ricochet	4.76	-2.13	8.05	10.46	-1.36	-0.90	-2.60	-2.59
Sampling without replacement								
The Da Vinci Code	12.94	12.29	12.87	12.95	-7.38	-5.58	-13.41	-12.13
The Five People you Meet in Heaven	11.05	10.80	10.88	10.97	-12.82	-10.27	-17.72	-30.38
The Rule of Four	13.14	12.76	11.02	14.08	-6.54	-5.09	-12.58	-15.30
R is for Ricochet	13.33	11.13	11.10	14.37	-3.85	-2.52	-6.65	-6.40

Notes:

Table 8: Supply side estimates: marginal costs and elasticities

To derive marginal costs we use equation (10). We replace store j's markets share q_j and own-price derivative $\partial q_j / \partial p_j$ by equations (7) and (9). What is left is an expression which only depends on search cost CDF and PDF values evaluated at the cutoff search costs as well as sampling probabilities, all of which have been reported above. Table 8 displays marginal costs for each bookstore-book combination as well as implied elasticities for both sampling assumptions. Table 9

[15] Amazon is sampled first in the majority of transactions (67%), followed by Barnes and Noble (16%), Book clubs (10%), and Other bookstores (7%).

gives markups (as a multiplier of marginal costs) and profit margins (in dollars) using the estimated elasticities reported in Table 8.

	Markups				Profit margins ($)			
Product name	Amazon	B&N	Book clubs	Other	Amazon	B&N	Book clubs	Other
Sampling with replacement								
The Da Vinci Code	1.277	1.428	1.112	1.129	3.248	4.486	1.406	1.612
The Five People you Meet in Heaven	1.138	1.186	1.083	1.045	1.453	1.881	0.879	0.484
The Rule of Four	1.831	1.267	1.268	1.207	7.039	9.444	2.532	2.587
R is for Ricochet	3.783	-8.670	1.624	1.628	13.242	20.578	5.018	6.569
Sampling without replacement								
The Da Vinci Code	1.157	1.218	1.081	1.090	2.028	2.684	1.037	1.163
The Five People you Meet in Heaven	1.085	1.108	1.060	1.034	0.934	1.165	0.651	0.373
The Rule of Four	1.181	1.245	1.086	1.070	2.372	3.120	0.952	0.984
R is for Ricochet	1.351	1.658	1.177	1.185	4.674	7.320	1.963	2.660

Notes:

Table 9: Supply side estimates: markups and margins

To derive marginal costs we use equation (10). We replace store j's markets share q_j and own-price derivative $\partial q_j/\partial p_j$ by equations (7) and (9). What is left is an expression which only depends on search cost CDF and PDF values evaluated at the cutoff search costs as well as sampling probabilities, all of which have been reported above. Table 8 displays marginal costs for each bookstore-book combination as well as implied elasticities for both sampling assumptions. Table 9 gives markups (as a multiplier of marginal costs) and profit margins (in dollars) using the estimated elasticities reported in Table 8. Table 10 shows the empirically observed market shares together with the market shares estimated using equation (7).

	Amazon		B&N		Book clubs		Other	
Product name	obs.	est.	obs.	est.	obs.	est.	obs.	est.
Sampling with replacement								
The Da Vinci Code	0.417	0.595	0.188	0.132	0.312	0.157	0.083	0.116
The Five People you Meet in Heaven	0.625	0.600	0.208	0.133	0.167	0.142	0.000	0.124
The Rule of Four	0.529	0.627	0.176	0.144	0.294	0.132	0.000	0.097
R is for Ricochet	0.348	0.607	0.130	0.138	0.522	0.148	0.000	0.107
Sampling without replacement								
The Da Vinci Code	0.417	0.568	0.188	0.123	0.312	0.178	0.083	0.131
The Five People you Meet in Heaven	0.625	0.576	0.208	0.123	0.167	0.157	0.000	0.143
The Rule of Four	0.529	0.612	0.176	0.138	0.294	0.143	0.000	0.107
R is for Ricochet	0.348	0.584	0.130	0.131	0.522	0.167	0.000	0.117

Notes:

Table 10: Observed and estimated market shares

7.1 Discussion

Our results on price elasticities and markups appear to depend on the sampling protocol (with replacement or without replacement). In general, we get much higher price elasticities, and lower markups for Amazon and Barnes and Noble from the model with sampling without replacement. This is perhaps intuitively clear – if sampling is with replacement, Amazon has higher market power in that the consumer will possibly sample it multiple times during her search. In future work, we will test whether actual searches satisfy the with or without replacement search protocol better. Obviously, product recall is a feature of non-sequential search, hence observing recall does not eliminate sampling without replacement. However, under sampling without replacement, we should not observe a consumer revisit more than one store a second time. A priori, however, we believe sampling without replacement is likely to be the more appropriate choice for this setting.

Our price elasticities also provide an interesting comparison with the results of Chevalier and Goolsbee (2003), who found own an own price elasticity of -3.5 for Barnes and Noble and -0.45 for Amazon, using the very different methodology of investigating the effect of price changes on sales ranks of books. Our estimated own price elasticity for Amazon are mostly a lot higher (between -1.3 and -12.8 across books and sampling protocols), and somewhat higher for Barnes and Noble (between -0.9 and -6.8). The difference between our findings may be due to several factors: first, Chevalier and Goolsbee's estimates are based on a much larger sample of books; our sample is restricted to four best-sellers. It is plausible that consumers are more price elastic when purchasing best-sellers (which could be utilized as "loss-leaders" by bookstores to attract new customers). Second, Chevalier and Goolsbee's results are based on 2001 data; whereas ours is based on 2004 data. It is possible that online bookshoppers in 2004 have gotten somewhat savvier at searching for deals than they were 2001. Third, our methodologies are quite different: while Chevalier and Goolsbee have the advantage of being able to utilize exogenous price shocks, but are limited by lack of sales/quantity data (and have to extrapolate using a Pareto distribution), our method relies crucially on the specification of our demand model. We hope that further research can identify data sets that can overcome the limitations of these two approaches.

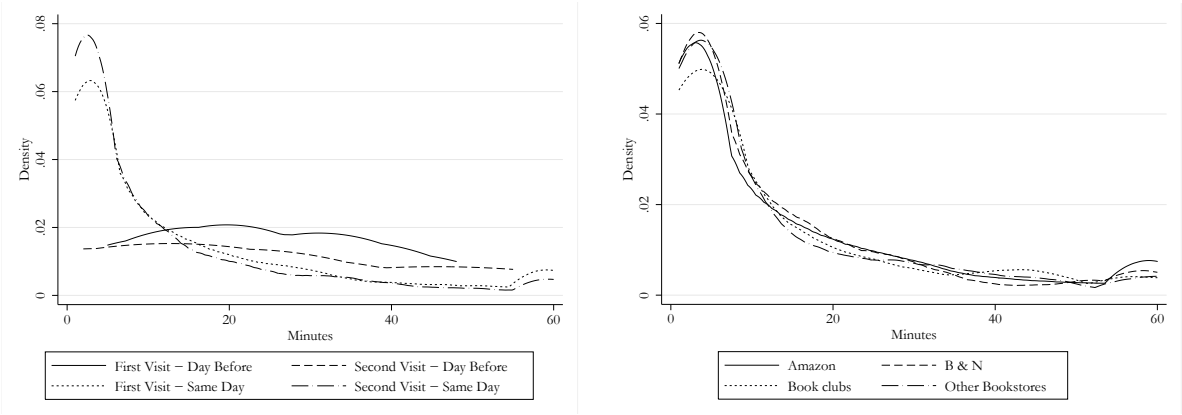
8 Conclusion

In this paper we have investigated to what extent consumers are indeed using the sequential and non-sequential search strategies put forth by the large theoretical literature on search behavior. By using detailed data on the browsing and purchasing behavior of a large panel of consumers, we have tested various restrictions classical search models put on search behavior. We have shown that the benchmark model of sequential search, where it is assumed consumers know the distribution of prices to sample from, can be rejected based on the recall patterns observed in the data, even if there is a finite number of firms. However, if consumers do not know the distributions from which prices are drawn but instead learn the price distributions using Bayesian updating, recall patterns no longer reject the sequential protocol. Instead, we have looked in more detail at patterns in the search costs that rationalize observed search behavior for given consumers over time, and shown using several tests that a non-sequential search model does a better job in explaining those patterns than a sequential search model with Bayesian updating.

Our finding that the non-sequential search protocol outperforms the sequential search model in terms of explaining observed search behavior for the subjects in our sample is to some extent surprising given that non-sequential search protocol is often thought of as a constrained version of sequential search. However, as shown by Morgan and Manning (1985) the optimal search model allows consumers to choose both the size of the sample and how many samples to take and as such encompasses both the sequential and non-sequential search protocol. When there is a large time lag between making the search decision and obtaining the actual quotation non-sequential search is typically optimal, because it allows the searcher to gather information quicker than would have been possible with sequential search.

Although a typical online shopper will not face large time lags when searching, a non-sequential search strategy might still be a good approximation of the optimal strategy if there exist economies of scale to sampling or if the searcher discounts the future. As argued by Manning and Morgan (1982), sufficiently large economies of scale to sampling will make it optimal to sample more firms at once and stop afterwards, even if the consumer can continue sampling. Indeed, after one has gone through the hassle of finding the right book and obtaining a price quote at one online bookstore, simple copying and pasting the ISBN number to the website of another bookstore is enough to

obtain an additional price quotation. Some preliminary evidence on whether this is indeed what is going on is presented in Figure 5(a), where we have plotted kernel density plots of the durations of the first and second search for searches on the same day and previous day of the transaction, conditional on searching more than once. As the graphs shows, searchers spend much less time during their second search. Figure 5(b) shows that this is not driven by the differences in the bookstores.



(a) Duration of First and Second Searches

(b) Duration of First Search per Firm

Figure 5: Duration Searches

Finally, we have explored the quantitative implications of our favored model, with non-sequential search, by estimating the price elasticities implied by the non-sequential search model, and the associated markups. Our findings indicate higher price elasticities than found in Chevalier and Goolsbee (2003), though in Section 7.1, we discuss several factors that may explain the differences in results. We hope that this exercise demonstrates the usefulness of the consumer search model as a "demand-side" model that could be applied in settings where consumer search is deemed an important factor.

APPENDIX

A: Probabilities of offering the lowest price

First consider sampling with replacement. With probability ρ_j the j-lowest ranked store is sampled first. With probability $(j/N)^{k-1}$ all remaining $k-1$ draws do not belong to stores offering lower prices than store j, so the probability of offering the lowest price out of k draws when being sampled first is $\rho_j(j/N)^{k-1}$. With probability $\rho_1 + \ldots + \rho_{j-1}$ a lower ranked store is sampled in first. In this case store j should at least be sampled once in the remaining draws. With probability $(j/N)^{k-1}$ no stores offering lower prices than store j will be drawn in the remaining draws. This probability includes combinations of stores that do not involve store j, i.e., with probability $(j-1)^{k-1}/N^{k-1}$ all $k-1$ draws will be stores offering lower utility than store j. Taking the difference gives the probability store j offers the lowest price among the remaining stores, i.e, this probability is $(j/N)^{k-1} - ((j-1)/N)^{k-1}$. Therefore, the probability of offering the lowest price out of k draws when not being sampled first is $(\rho_1 + \ldots + \rho_{j-1})\left((j/N)^{k-1} - ((j-1)/N)^{k-1}\right)$. Taken together, the probability the j-lowest ranked firm offers the lowest price out of $k = 3$ random draws or more, is

$$\alpha_{jk} = \rho_j \left(\frac{j}{N}\right)^{k-1} + (\rho_1 + \ldots + \rho_{j-1}) \left(\left(\frac{j}{N}\right)^{k-1} - \left(\frac{j-1}{N}\right)^{k-1} \right).$$

Next consider sampling without replacement. When sampling k times, k firms need to be picked out of N firms. With probability ρ_j the j-lowest ranked firm is the starting point. Out of the remaining $N-1$ firms, $k-1$ firms need to be picked, which all have to offer a higher price than firm j in order for firm j to offer the lowest price. There are $j-1$ such stores, so the probability that store j sells conditional on being the first sampled can be calculated using the hypergeometric distribution, i.e., this probability is $\binom{j-1}{k-1}/\binom{N-1}{k-1}$. With probability $\rho_1 + \ldots + \rho_{j-1}$ one of the other stores is the starting point for the consumer. In that case store j has to be sampled in one of the remaining searches, which is proportional to $k-1$. The remaining $k-2$ stores sampled need to offer higher prices; to calculate this probability we can again use the hypergeometric distribution, i.e., the probability is $\binom{j-2}{k-2}/\binom{N-2}{k-2}$. All together, the probability the j-lowest ranked firm offers

the lowest price out of $k = 3$ random draws or more, where $j \geq k$, is

$$\begin{aligned}
\alpha_{jk} &= \rho_j \frac{\binom{j-1}{k-1}}{\binom{N-1}{k-1}} + (\rho_1 + \ldots + \rho_{j-1}) \frac{k-1}{N-1} \frac{\binom{j-2}{k-2}}{\binom{N-2}{k-2}}; \\
&= \rho_j \frac{(N-k)!(j-1)!}{(j-k)!(N-1)!} + (\rho_1 + \ldots + \rho_{j-1}) \frac{k-1}{N-1} \frac{(N-k)!(j-2)!}{(j-k)!(N-2)!}; \\
&= \left((\rho_1 + \ldots + \rho_{j-1}) \frac{k-1}{j-1} + \rho_j \right) \frac{(N-k)!(j-1)!}{(j-k)!(N-1)!}; \\
&= \left((\rho_1 + \ldots + \rho_{j-1}) \frac{k-1}{j-1} + \rho_j \right) \frac{(j-1) \times \ldots \times (j-(k-1))}{(N-1) \times \ldots \times (N-(k-1))}.
\end{aligned}$$

When $j < k$ store j will never offer the lowest price, so $\alpha_{jk} = 0$ if $j < k$.

B: Derivatives of Demand Curves

Using equations $(6a)$ and $(6b)$, first rewrite the market share equation (7) as

$$q_j = \alpha_{j1} - \sum_{k=1}^{N-1} (\alpha_{jk} - \alpha_{j(k+1)}) G(c_k).$$

Taking the price derivative gives

$$\frac{dq_j}{dp_j} = - \sum_{k=1}^{N-1} (\alpha_{jk)} - \alpha_{j(k+1)}) g(c_k) \frac{dc_k}{dp_j} \tag{A11}$$

The derivative of c_k with respect to p_j is

$$\frac{dc_k}{dp_j} = \alpha_{jk} - \alpha_{j(k+1)}.$$

Plugging this in equation (A11) gives

$$\frac{\partial q_j}{\partial p_j} = - \sum_{k=1}^{N-1} (\alpha_{jk} - \alpha_{j(k+1)})^2 g(c_k).$$

References

[1] Bo Axell: "Search market equilibrium," *Scandinavian Journal of Economics* 79, 2040, 1977.

[2] Colin Camerer: "Individual decision making," in *Handbook of Experimental Economics*, edited by John H. Kagel and Alvin E. Roth, Princeton University Press, Princeton, NJ, 587-703, 1995.

[3] John A. Carlson and R. Preston McAfee: "Discrete equilibrium price dispersion," *Journal of Political Economy* 91, 480-93, 1983.

[4] Xiaohong Chen, Han Hong, and Matthew Shum: "Nonparametric likelihood ratio model section tests between parametric likelihood and moment condition models," *Journal of Econometrics* 141, 109-40, 2007.

[5] Judith Chevalier and Austan Goolsbee: "Measuring prices and price competition online: Amazon and Barnes and Noble," *Quantitative Marketing and Economics* 2, June 2003

[6] Chien-fu Chou and Gabriel Talmain: "Nonparametric search," *Journal of Economic Dynamics and Control* 17, 771-84, 1993.

[7] Morris H. DeGroot: "Some problems of optimal stopping," *Journal of the Royal Statistical Society. Series B (Methodological)* 30, 108-22, 1968.

[8] Morris H. DeGroot: "Optimal Statistical Decisions," McGraw-Hill, New York, 1970.

[9] Babur I. De los Santos: "Consumer Search on the Internet," Unpublished Manuscript, 2008.

[10] Glenn W. Harrison and Peter Morgan: "Search intensity in experiments," *Economic Journal* 100, 478-86, 1990.

[11] Han Hong and Matthew Shum: "Using price distributions to estimate search costs," *RAND Journal of Economics* 37, 257-75, 2006.

[12] Carl A. Kogut: "Consumer search behavior and sunk costs," *Journal of Economic Behavior and Organization* 14, 381-92, 1990.

[13] Richard Manning and Peter Morgan: "Search and Consumer Theory," *Review of Economic Studies* 49, 203-16, 1982.

[14] John J. McCall: "Economics of Information and Job Search," *Quarterly Journal of Economics* 84, 113-26, 1970.

[15] José Luis Moraga-González and Matthijs R. Wildenbeest: "Maximum Likelihood Estimation of Search Costs," *European Economic Review* 52, 820-48, 2008.

[16] Peter Morgan and Richard Manning: "Optimal Search," *Econometrica* 53, 923-55, 1985.

[17] Dale T. Mortensen: "Job search, the duration of unemployment and the Phillips curve," *American Economic Review* , 84762, 1970.

[18] Jennifer F. Reinganum: "A Simple Model of Equilibrium Price Dispersion," *Journal of Political Economy* 87, 851-58, 1979.

[19] Donald B. Rosenfield and Roy D. Shapiro: "Optimal adaptive price search," *Journal of Economic Theory* 25, 1-20, 1981.

[20] Michael Rothschild: "Searching for the lowest price when the distribution of prices is unknown," *Journal of Political Economy* 82, 689-711, 1974.

[21] Andrew Schotter and Yale M. Braunstein: "Economic search: an experimental study," *Economic Inquiry* 19, 1-25, 1981.

[22] Joep Sonnemans: "Strategies of search," *Journal of Economic Behavior and Organizaton* 35, 309-32, 1998.

[23] Martin L. Weitzman: "Optimal search for the best alternative," *Econometrica* 47, 641-54, 1979.

[24] Rami Zwick, Amnon Rapoport, Alison King Chung Lo, and A. V. Muthukrishnan: "Consumer sequential search: not enough or too much?" *Marketing Science* 22, 503-19, 2003.