

TESTING MODELS OF CONSUMER SEARCH USING DATA ON WEB BROWSING AND PURCHASING BEHAVIOR *

Babur De los Santos[†]
Ali Hortacısu[‡]
Matthijs R. Wildenbeest[§]

First version: April 2009
This version: December 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 both the recall patterns we observe in the data and the absence of dependence of search decisions on observed prices. 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 fixed sample size search than sequential search with learning. Our findings suggest fixed sample size search provides a more accurate description of observed consumer search behavior. We then utilize the fixed sample size search model to estimate the price elasticities and profit margins 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 *fixed sample size* 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 fixed sample size 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 fixed sample size 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 fixed sample size 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 fixed sample size 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

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

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

³See 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 fixed sample size 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 nonparametrically estimated fixed sample size 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 fixed sample size 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 fixed sample size 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 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 two variants of the sequential search model in which consumers might find it optimal to buy from a previously visited firm. In the first variant consumers believe utilities are sampled from alternative specific utility distribution, while in the second variant the assumption that consumers “know” the distribution of utilities while deciding on their search strategy is relaxed. Importantly, in these settings, the sequential search model can not be rejected based on recall patterns alone. Instead, we look at other testable predictions of these models. For instance, the sequential search model with utilities drawn from independent probability distributions implies consumers should start at the alternative with the highest reservation utility. In Section 4 we find that this prediction does not hold for a large number of consumers. We also do not observe any

dependence of the decision to continue searching on observed prices, which one would expect if consumers would be searching according to one of the sequential search variants.

In Section 5 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 this section we also 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 fixed sample size search assumption display much less dispersion within person than the search costs estimated under the sequential search with Bayesian learning model. This means the fixed sample size search model leads to more stable parameter estimates, and we thus conclude that fixed sample size search may provide a more accurate description of observed behavior.

Finally, in Section 7 we use the favored fixed sample size 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 fixed sample size 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

⁴Although 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.

⁵Each observation represents a single book purchased during one transaction; if multiple copies of the book are

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

Bookstore	Transactions		Visits	
	Number	%	Number	%
Amazon	10,206	65.5%	249,593	76.3%
Barnes and Noble	3,046	19.6%	25,758	7.9%
<i>Book Clubs</i>				
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%
<i>Other Bookstore</i>				
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 purchased in the same transaction, it is recorded as one observation.

⁶This 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.

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 the transaction. Table 9 gives descriptive statistics of the sample.

	2002		2004	
	Mean	Std. Dev.	Mean	Std. Dev.
<i>Duration of each website visit (in minutes)</i>				
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 Transactions only	19.04	18.26	15.74	17.37
	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

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

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.

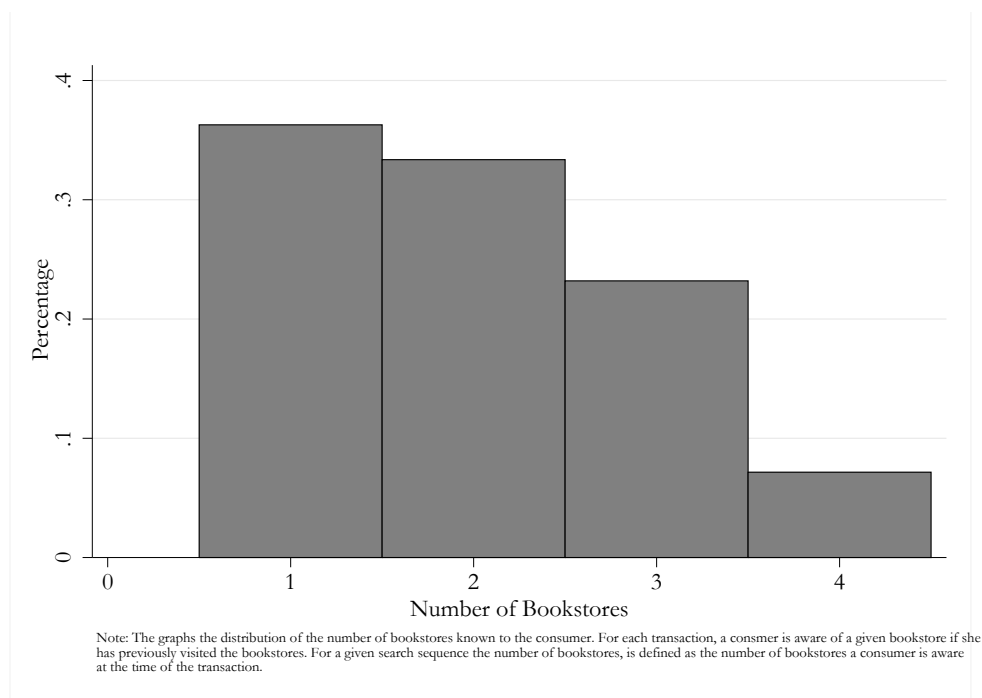


Figure 1: Consumer Bookstore Awareness

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. 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 forty-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 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.⁸ 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 <i>max</i> and <i>min</i>	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, \bar{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

We study a setting in which consumers inelastically demand one unit of a good sold by a finite number of stores. Before searching consumers do not know the realized utility levels of the available alternatives; in order to have this revealed a consumer has to pay a constant search cost c per alternative, which is assumed to be randomly drawn from a search cost distribution. We study two

⁸We 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.

approaches in modeling consumers' search behavior: sequential and fixed sample size search.

Consider first the fixed sample size search paradigm. Assuming the consumer believes that each store's utility is an i.i.d. draw from distribution $F(u)$ with density $f(u)$ and support between \underline{u} and \bar{u} , a consumer will determine the optimal number of stores k in her sample by maximizing expected utility net of total search cost:

$$k(c) = \arg \max_k \int_{\underline{u}}^{\bar{u}} k \cdot u F(u)^{k-1} f(u) du - k \cdot c.$$

The integral on the right-hand side is the expected maximum utility drawn from the utility distribution when sampling k alternatives, so the optimal sample size is determined by finding sample size k that maximizes the difference between the expected maximum utility out of searching k times and the total search cost $k \cdot c$.

Next consider sequential search. As shown by McCall (1970) a consumer will continue to search as long as she finds a utility lower than some reservation value $r(c)$, where $r(c)$ is given by:

$$c = \int_{r(c)}^{\bar{u}} (u - r(c)) f(u) du. \tag{1}$$

As seen in the equation, the reservation value is such that, if the utility 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 increase in utility from another search, accounting for the option value of discarding lower utility draws.

One important difference between the two search paradigms is that fixed sample size search allows for buying from previously visited alternatives while sequential search does not. By definition fixed sample size search implies a consumer first samples all alternatives in a given sample and then decides which alternative to purchase. Therefore, if a consumers searches multiple times and the alternative sampled first offers the best deal a consumer will return there to buy. However, for a consumer searching sequentially at a search cost c the reservation value $r(c)$ is constant across searches, which means the consumer will never recall an alternative that she sampled earlier, unless there are a finite number of stores, and the consumer has visited all the stores. Our first test will focus on recall behavior by consumers.

Test 1 (No Recall) *Under the null hypothesis of the standard sequential search 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.*

A second difference between the sequential and fixed sample size paradigms is that the optimal sample size under fixed sample size is independent of observed utilities, while under sequential search the decision to continue searching or not should depend on observed utility values. Assuming price is a major determinant of utility, consumers searching sequentially are therefore more likely to continue searching when a relatively high priced is observed. This also means that consumers who search only once are more likely to have encountered a relatively low price. Our second test focuses on this dependence of search decisions on observed prices.

Test 2 (Price Dependence) *Under the null hypothesis of the standard sequential search model, consumers searching once are more likely to have found a relatively low price, while the first price observation of consumers searching twice is likely to be relatively high.*

Observe that the absence of recall in the sequential model described above depends crucially on the constant reservation value rule. We now discuss two variants of the classic sequential search model that may lead to a sequence of reservation values that are decreasing in the number of alternatives sampled. The first variant is a setting where the utility of each alternative has its own independent probability distribution. As shown by Weitzman (1979) in this case the optimal procedure is to start searching at the alternative with the highest reservation utility and to terminate search whenever the maximum sampled utility exceeds the reservation utilities of all remaining unsampled alternatives. After each search the reservation utilities of the unsampled alternatives will go down, which means that a previously sampled alternative might not pass the threshold initially but do so later on, resulting in recall. Our second test will focus on the optimal starting point in case of independent probability distributions.

Test 3 (Search Order) *Under the null hypothesis of a standard sequential search model with utilities drawn from independent probability distributions, consumers should start at the alternative with the highest reservation utility.*

A second variant of the classical sequential search model that may lead to decreasing reservation values in the number of sampled alternatives is a setting where consumers update their beliefs about the utility distribution $F(u)$ while sampling. As shown by Rosenfield and Shapiro (1981), if searchers learn by Bayesian updating Dirichlet priors over utility levels that follow a multinomial

distribution consumers's optimal search policy is myopic and can be characterized by a reservation value that is non-increasing in the number of alternatives sampled. More specifically, suppose a random vector of utilities u_1, u_2, \dots, u_n , where $u_j > u_{j+1}$, follows a multinomial distribution with probabilities of each utility $\pi_1, \pi_2, \dots, \pi_n \geq 0$ and $\sum_{j=1}^n \pi_j = 1$.⁹ The true probability of each utility being sampled is unknown, but its prior distribution is assumed to be Dirichlet with parameters $\alpha_1, \alpha_2, \dots, \alpha_n$, where $\alpha_j / \sum_k \alpha_k$ can be interpreted as consumers' believe of the probability of observing utility u_j . Since the Dirichlet is the conjugate prior of the multinomial distribution the posterior is Dirichlet as well. This means the posterior distribution after having observed an alternative with utility u_j is updated to $\alpha_1, \alpha_2, \dots, \alpha_j + 1, \dots, \alpha_n$. Moreover, in case consumers start with an (uninformative) uniform Dirichlet prior with weight $W = \sum_k \alpha_k$, the expected probability of finding a better deal than the best deal observed so far, denoted by u^{\max} , is given by $(\bar{u} - u^{\max}) \cdot W / (W + t)$, where t is the number of sampled alternatives so far. Since the expected gain in utility conditional on finding a better deal is $(\bar{u} - u^{\max}) / 2$, the expected gains from search G at u^{\max} are $G(u^{\max}) = (1/2) \cdot (\bar{u} - u^{\max})^2 \cdot W / (W + t)$. The reservation value is defined as the search cost that makes a consumer indifferent between searching once more and consuming the alternative at hand, so $c = (1/2) \cdot (\bar{u} - r(c, t))^2 \cdot W / (W + t)$, or

$$r(c, t) = \bar{u} - \sqrt{2c \cdot \frac{W + t}{W}}.$$

Since the reservation value $r(c, t)$ is decreasing in the number of alternatives sampled this means that also a sequential search model with Bayesian updating can explain why a consumer may return to previously visited stores, even if they have not exhausted all their search possibilities. Test 1 thus cannot rule out a sequential search model with learning. Notice however that Test 2 also applies to the model of sequential search with updating— consumers encountering a relatively high price at their first visit are more likely continue searching, while consumers who search once should have encountered a relatively low price, even if consumers update their initial priors. In addition, although the model allows for recall, consumers will only do so under specific circumstances: if a consumers with search cost c returns to a previously sampled alternative to purchase this means $r(c, t + 1) \leq u^{\max} < r(c, t)$, or

$$\frac{W + t}{W + t + 1} G(u^{\max}) \leq c < G(u^{\max}).$$

⁹The case of a continuous distribution is studied by Bikhchandani and Sharma (1996) and leads to similar results. See also Häubl, Dellaert, and Donkers (2009).

If a consumer returns to a previously sampled alternative to buy, that is, the consumer recalls, the range of search costs that rationalizes this behavior is thus relatively small. This means if we observe consumers recalling, given the support of the utility 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 Bayesian updating can explain the recall patterns we observe.

4 Test results

In this section we use the data on consumer browsing behavior to see whether it is in line with the empirical implications of several consumers search models, using the tests developed in the previous section.

4.1 Testing the “no recall” hypothesis

The benchmark sequential search model 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 sampled alternative, which means this alternative should have a utility below the reservation value.

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. For instance, if we observe that the consumer has only visited Amazon and Barnes & Noble 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 fixed sample size 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 percentage 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.¹⁰ Moreover, Table 5 also shows that Amazon 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.

4.2 Testing the “price dependence” hypothesis

The results of the previous sections rule out the basic sequential search model with a constant reservation value strategy. However, as we argued above, once we allow for Bayesian updating or for alternative specific utility distributions the observation of recall no longer invalidates sequential search. A second implication of sequential search is that the decision to continue searching depends on the prices observed, even in a model with Bayesian updating. To test this hypothesis we check whether (i) consumers searching once have found a lower price than expected based on the overall price distribution, and (ii) whether consumers searching twice have found relatively higher price

¹⁰Note 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.

on their first visit than consumers who have decided to search once. Note that for less than one percent of transactions we have prices of all four bookstores, so for these tests we focus on Amazon and Barnes & Noble only.

Buy from	Lowest price			# cons
	Amazon	equal	B&N	
Amazon	45.63%	23.18%	31.20%	686
Barnes & Noble	49.87%	18.35%	31.78%	387
total	47.16%	21.44%	31.41%	1,073

Notes:

Table 6: Price first store

Table 6 shows that consumers who search once and buy from Amazon in 46% of transactions have found the lowest price. However, this percentage is lower than the percentage of all transactions in which Amazon has the lowest price, which suggests that these consumers as a group have not found better prices than those randomly drawn from the overall price distribution. Similarly, consumers who search once and buy from Barnes & Noble do not get better prices on average than the Barnes & Noble price one could expect when searching once. These findings show that consumers do not seem to exercise their continue-to-search option when necessary, which violates the sequential search protocol.

Price first store	once	twice	total
lowest or equal	62.07%	63.87%	62.30%
highest	37.93%	36.13%	37.70%
# consumers	1,073	155	1,228

Notes:

Table 7: Price first store by number of searches

To see whether consumers searching once have found on average lower prices on their first visit than consumers searching twice, Table 7 gives the percentage of consumers having found the lowest or equally lowest price at the store visited first for consumers searching once and for consumers searching twice. For both groups of consumers the price at the first store is the lowest or equal in approximately the same percentage of transactions: in 62% of the transactions in which consumers search once the store visited first has the lowest price, while for transaction in which consumers search twice this percentage is 64%. This suggests consumers searching twice have not found worse

prices at the first store visited than consumers searching once. If consumers would be searching sequentially one would expect to observe at least some dependence.

4.3 Testing the “optimal starting point” hypothesis

As summarized by Test 3 optimal sequential search for alternatives with utilities drawn from independent probability distribution implies consumers should start searching at the store that gives the highest expected utility. To test this hypothesis we check whether consumers who have engaged in multiple transactions and during the sampling period have always bought from the same store are also beginning their search sequence at that specific store. For almost sixty percent of the 24% of consumers who can be qualified as “loyal” to one of the stores this is not the case: only 41% start at their preferred store. To see whether there are differences across the four stores we split up the store visited first by revealed store preference in Table 8. The table shows that around 52% of Amazon’s transactions involving consumers buying only from Amazon visit Amazon first, which suggests the majority of Amazon’s loyal consumers indeed start at their preferred outlet. However, as shown in the other columns of Table 8 an even higher percentage of consumers having revealed a preference for one of the other stores also visit Amazon first. Altogether this suggests that independent of store preference, most consumers prefer to start at Amazon, which rules out a sequential search model where consumers optimally start at their preferred store. Moreover, the observation that the majority of consumers loyal to one of Amazon’s competitors do start at Amazon when searching more than once suggests that these consumers already anticipate searching more than once, contradicting the sequential search protocol.

Start at	Loyal to					Total
	Amazon	B&N	Bookclubs	Other	None	
Amazon	52.44%	54.22%	56.16%	61.22%	55.24%	54.78%
Barnes & Noble	17.99%	26.62%	8.22%	8.16%	17.89%	18.12%
Bookclubs	7.01%	6.82%	21.23%	2.04%	8.64%	8.61%
Other	22.56%	12.34%	14.38%	28.57%	18.23%	18.49%
# consumers	656	308	146	49	2,616	3,775

Notes:

Table 8: Starting store consumers searching more than once by loyalty

5 Estimates of search costs implied by sequential and fixed sample size search models

Although the price-dependence test results do not provide support for both sequential search with and without learning in this section we proceed with an alternative testing strategy that focuses on some other aspects of the sequential search model with learning: we will estimate search cost bounds implied by both fixed sample size 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. Unfortunately our dataset is not rich enough to estimate a model where firms are heterogenous in other dimensions than price. Nevertheless, to capture our finding above that 24% of consumers engage in multiple transactions but only buy from one store and thus seem to have a specific store preference independent of price, we also estimate search costs for those consumers that are not found to be loyal to any of the stores.

5.1 Bounds generated by sequential search with Bayesian updating

Recall from Section 3 that under a sequential search model with Bayesian updating, if we observe a consumer searching twice but buying from the first store we know her search cost c should have been

$$\frac{W+t}{W+t+1}G(p^{\min}) \leq c < G(p^{\min}),$$

where, assuming a uniform Dirichlet prior, the gains from search after having observed price p^{\min} are

$$G(p^{\min}) = (1/2) \cdot (p^{\min} - \underline{p})^2 \cdot \frac{W}{W+t}.$$

Note that these equations are now written in terms of prices, assuming stores are similar in other aspects. To estimate the model we need consumers' weight put on prior W , the lower bound of the price distribution \underline{p} and the transaction price p^{\min} . We estimate W by the number of stores 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.

How precise our estimates of search costs are going to be depends on the observed search patterns. For transactions in which consumers buy after having searched once we can only infer

the lower bound of search cost values that can rationalize this behavior, i.e.,

$$\underline{c}^s = G(p^{\min}).$$

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

$$\underline{c}^s = \frac{W + t}{W + t + 1} G(p^{\min}).$$

As shown above, the upper bound of our search cost estimate for those who searched more than one firm and recalled is

$$\bar{c}^s = G(p^{\min}).$$

For people who searched more than one firm and bought from the last one, all we know is that the first observed price must have been lower than the upper bound of the price distribution, which means that the upper bound on the implied search cost is equal to the gains from search at the upper bound, i.e.,

$$\bar{c}^s = G(\bar{p}) = \frac{\bar{p} - \underline{p}}{2}.$$

5.2 Bounds generated by a fixed sample search search model with uniform distribution

To be able to compare search cost estimates of the sequential search model with Bayesian updating we also estimate search costs using a fixed sample search model. First define $c(k)$ as the search cost of a consumer searching k times, that is, $c(k) = Ep_{1:k} - Ep_{1:k+1}$, where $Ep_{1:k}$ denotes the expected lowest price out of k searches. Assuming consumers expect the price distribution to be uniform between \underline{p} and \bar{p} , the search cost cutoff values are given by

$$c(k) = \frac{\bar{p} - \underline{p}}{k^2 + 3k + 2}.$$

For a transaction the estimated bounds are therefore

$$\underline{c}^{ns} = \frac{\bar{p} - \underline{p}}{k^2 + 3k + 2} \leq c \leq \frac{\bar{p} - \underline{p}}{k^2 + k} = \bar{c}^{ns}.$$

To estimate these bounds we only have to observe the bounds of the price distribution and the number of searches. Note that similar to sequential search we do not observe the upper bound for those who only visit one firm.

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.¹¹ We then calculate the within consumer standard deviation of the “midpoint” search cost estimates. Figure 2(a) displays the empirical CDF of the within-consumer standard deviation of search costs implied by the sequential and fixed sample size search models.¹² For both sequential and fixed sample size search the sample consists of consumers who have searched at least twice multiple times, so we can calculate a consumer’s standard deviation of search costs across transactions. The gray curves depict our estimates conditional on having bought at different stores during the sampling period, while the blue (sequential) and black (fixed sample size) curve also includes the consumers who are loyal to one of the stores— clearly there is not much difference.

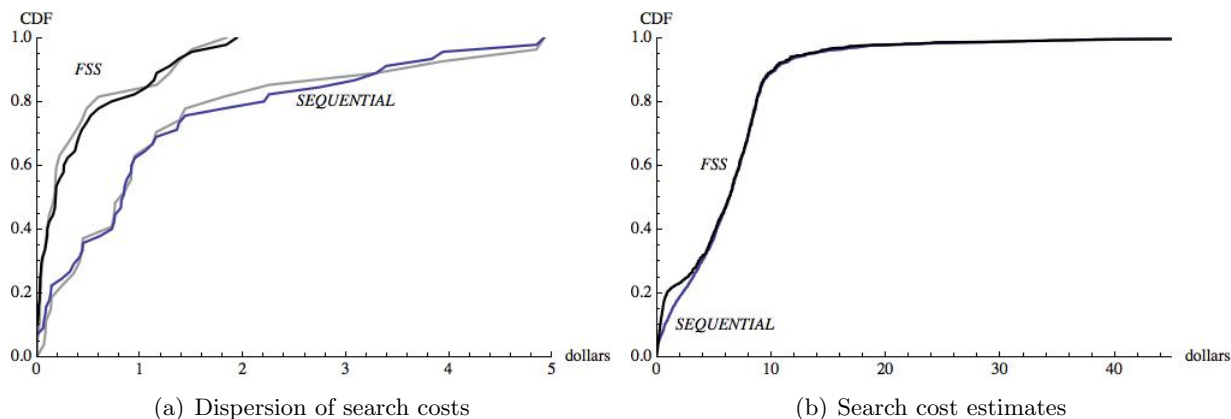


Figure 2: Results bound estimates

Observe that our estimates of search costs based on the fixed sample size model display smaller within-person dispersion than our estimates based on the sequential model. If we believe search costs to be relatively time invariant, the figure suggests that the fixed sample size model does a better job explaining our data.

Although the fixed sample size search model appears to provide estimates that are more consistent across consumers engaging in multiple transactions, Figure 2(b) suggests that our inference

¹¹Since 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.

¹²Although the “midpoint” is a somewhat arbitrary summary of the bounds, the figure does not change qualitatively if we plot the standard deviations of the bounds separately.

regarding the distribution of search costs might not have been very sensitive to the choice of modeling paradigm. In this figure we display the empirical distribution of the search cost estimates across transaction we obtained from the two models, but this time we also include search cost estimates for consumers who search once. To be able to do so we assume a consumer will only participate in the market if the expected gains from searching once outweigh the cost of searching once. These gains are the same for both search protocols, which means that our estimate of the search cost upper bound for these consumers is $\bar{c}^s = \bar{c}^{ns} = (\bar{p} - \underline{p})/2$. Since approximately 75% of transactions involve consumers searching once, our finding that the estimated search cost distributions overlap for the upper three quartiles of the distribution is not very surprising. The main divergence between the model occurs at the lower end of the distributions, corresponding to transactions of consumers who search more than once, with the fixed sample size model implying more mass at zero.

6 Implications of the fixed sample size search model

We will now investigate price elasticities and (static) profit-maximizing firm behavior in an environment where consumers search using a fixed sample size search strategy. Based on the patterns we observed in our data, we allow for unequal *first* sampling probabilities as well as marginal cost heterogeneity. Unfortunately our data is not rich enough to be able to fully control for consumer heterogeneity in store preferences. However, in line with our findings above we allow for a share λ of consumers to be loyal to one of the bookstores— we assume the remaining $1 - \lambda$ consumers are searching for the lowest price.

On the supply side 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 the cost of sampling a price observation is c , where c assumed to be randomly drawn from a search cost distribution $G(c)$.

We assume the share of $1 - \lambda$ non-loyal consumers search using a fixed sample size strategy and have 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 one of the other bookstores.

The assumption that consumers observe the empirical price distribution function allows us to label the N stores by descending prices, $p_1 > \dots > p_N$, 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 $\alpha_{j2} = (\rho_1 + \dots + \rho_{j-1} + j\rho_j)/N$. 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.,

$$\alpha_{jk} = \rho_j \left(\frac{j}{N}\right)^{k-1} + (\rho_1 + \dots + \rho_{j-1}) \left(\left(\frac{j}{N}\right)^{k-1} - \left(\frac{j-1}{N}\right)^{k-1} \right). \quad (2)$$

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 $\alpha_{j2} = (\rho_1 + \dots + \rho_{j-1} + (j-1)\rho_j)/(N-1)$. 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.,

$$\alpha_{jk} = \begin{cases} \left((\rho_1 + \dots + \rho_{j-1}) \frac{k-1}{j-1} + \rho_j \right) \frac{(j-1) \times \dots \times (j-(k-1))}{(N-1) \times \dots \times (N-(k-1))} & \text{if } j \geq k; \\ 0 & \text{if } j < k. \end{cases} \quad (3)$$

Equations (2) or (3) can be used to characterize optimal consumer behavior as well the supply side of the market. Consider first the consumer side of the market. Non-loyal consumers are characterized by a search cost value which is drawn from a search cost distribution $G(c)$ with density function $g(c)$. The fixed sample size 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

$$c_k = \sum_{j=1}^N (\alpha_{jk} - \alpha_{j(k+1)}) p_j. \quad (4)$$

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.,

$$\mu_1 = 1 - G(c_1) \text{ for } k = 1; \quad (5a)$$

$$\mu_k = G(c_{k-1}) - G(c_k) \text{ for } k = 2, 3, \dots, N, \quad (5b)$$

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 (2) or (3) and the grouping of consumers μ_k given in equations (5a) and (5b) to calculate the market shares, i.e., the market share equation for store j is just its share of loyal consumers, denoted λ_j , where $\sum_j \lambda_j = \lambda$, plus the sum of the probability of selling to the different groups of non-loyal consumers, multiplied by their shares in the consumer population:

$$q_j = (1 - \lambda) \sum_{k=1}^N \alpha_{jk} \mu_k + \lambda_j. \quad (6)$$

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.,

$$q_j + (p_j - mc_j) \frac{\partial q_j}{\partial p_j} = 0. \quad (7)$$

In the Appendix we show the derivatives of the market share equations (6) are

$$\frac{\partial q_j}{\partial p_j} = -(1 - \lambda) \sum_{k=1}^{N-1} (\alpha_{jk} - \alpha_{j(k+1)})^2 g(c_k). \quad (8)$$

6.1 Estimation

We observe overall sampling probabilities and product specific prices, so we can directly calculate the c_k 's defined in equation (4). We also observe μ_k , the shares of consumers searching k times, from which we can calculate $G(c_k)$ for $k = 1, 2, \dots, N$ using equations (5a) and (5b). 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 (6). Equation (8) can be used to estimate the derivatives 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 profit margins as

$$p_j - mc_j = -\frac{q_j}{\partial q_j / \partial p_j}. \quad (9)$$

7 Application: price elasticities and profit margins of online bookstores

In this section, we present estimates price elasticities and profit margins 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.¹³ 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.

Product name	Obs.	Prices (\$)				Consumers by Sample Size (%)		
		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
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

Table 9: Descriptive Statistics

Table 9 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.

Product name	Cutoff search costs			CDF values			PDF values		
	c_1	c_2	c_3	$G(c_1)$	$G(c_2)$	$G(c_3)$	$g(c_1)$	$g(c_2)$	$g(c_3)$
<i>Sampling with replacement</i>									
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
<i>Sampling without replacement</i>									
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	1.021	0.836	0.710	0.154	0.000	0.000	1.662	0.000	0.000
R is for Ricochet	1.228	0.884	0.614	0.198	0.036	0.000	0.672	0.268	0.000

Table 10: Empirical Non-Sequential Search Cost CDF

Table 10 gives the estimated cutoff values of the search cost distributions that rationalize observed search patterns. These cutoff search costs are estimated using equation (4). We report our findings for sampling with replacement as well sampling without replacement. We allow for asymmetric first sampling probabilities – the probability a bookstore is sampled first is estimated using

¹³Book clubs did not appear on mySimon.com during the sampling period, so for this category we use sales weighted median transaction prices.

all transactions in the database.¹⁴ 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 10.

Product name	Profit margins (\$)				Elasticities E			
	Amazon	B&N	Book clubs	Other	Amazon	B&N	Book clubs	Other
<i>Sampling with replacement</i>								
The Da Vinci Code	6.06	7.95	2.07	1.81	-2.47	-1.88	-6.74	-7.79
The Five People you Meet in Heaven	2.70	3.32	1.33	0.54	-4.44	-3.61	-8.64	-23.02
The Rule of Four	12.82	16.12	3.94	2.97	-1.21	-0.99	-3.03	-5.07
R is for Ricochet	24.46	35.79	7.52	7.45	-0.74	-0.52	-1.74	-2.29
Mean	11.51	15.79	3.71	3.19	-2.21	-1.75	-5.04	-9.04
<i>Sampling without replacement</i>								
The Da Vinci Code	3.87	4.92	1.46	1.29	-3.87	-3.04	-9.50	-10.94
The Five People you Meet in Heaven	1.77	2.13	0.96	0.41	-6.77	-5.62	-12.07	-27.63
The Rule of Four	4.37	5.43	1.44	1.12	-3.55	-2.93	-8.82	-13.50
R is for Ricochet	8.79	13.02	2.83	2.98	-2.05	-1.41	-5.31	-5.71
Mean	4.70	6.47	1.67	1.45	-4.06	-3.25	-8.92	-14.44

Table 11: Supply side estimates: profit margins and elasticities

To derive profit margins we use equation (9). We replace store j 's markets share q_j and own-price derivative $\partial q_j / \partial p_j$ by equations (6) and (8). 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 11 displays profit margins (in dollars) for each bookstore-book combination as well as implied elasticities for both sampling assumptions.

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 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 fixed sample size search, hence observing recall does not

¹⁴Amazon is sampled first in the majority of transactions (67%), followed by Barnes and Noble (16%), Book clubs (10%), and Other bookstores (7%).

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 higher (between -0.7 and -6.8 across books and sampling protocols), but a bit lower for Barnes and Noble (between -0.5 and -5.6). 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 book shoppers 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 fixed sample size 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. In addition we do not find support for any price dependence of search decisions— if consumers would be searching sequentially they should be more likely to continue searching when a relatively high price is observed.

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 fixed sample size search model does a better job in explaining those patterns than a sequential search model with Bayesian updating.

Our finding that the fixed sample size 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 fixed sample size 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 fixed sample size search protocol. When there is a large time lag between making the search decision and obtaining the actual quotation fixed sample size 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 fixed sample size 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.

Finally, we have explored the quantitative implications of our favored model, with fixed sample size search, by estimating the price elasticities implied by the fixed sample size 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 $\rho_1 + \dots + \rho_{j-1}$ 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 $(\rho_1 + \dots + \rho_{j-1}) \left((j/N)^{k-1} - ((j-1)/N)^{k-1} \right)$. Taken together, the probability the j -lowest ranked firm offers the lowest price out of $k=3$ random draws or more, is

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

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 $\rho_1 + \dots + \rho_{j-1}$ 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

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

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 (5a) and (5b), first rewrite the market share equation (6) as

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

Taking the price derivative gives

$$\frac{dq_j}{dp_j} = -(1 - \lambda) \sum_{k=1}^{N-1} (\alpha_{jk} - \alpha_{j(k+1)}) g(c_k) \frac{dc_k}{dp_j} \quad (\text{A10})$$

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 (A10) gives

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

References

- [1] Bo Axell: “Search market equilibrium,” *Scandinavian Journal of Economics* 79, 2040, 1977.
- [2] Sushil Bikhchandani and Sunil Sharma: “Optimal search with learning,” *Journal of Economic Dynamics and Control* 20, 333-59, 1996.
- [3] 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.
- [4] John A. Carlson and R. Preston McAfee: “Discrete equilibrium price dispersion,” *Journal of Political Economy* 91, 480-93, 1983.
- [5] 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.
- [6] Judith Chevalier and Austan Goolsbee: “Measuring prices and price competition online: Amazon.com and BarnesandNoble.com,” *Quantitative Marketing and Economics* 1, 203-22, 2003.
- [7] Chien-fu Chou and Gabriel Talmain: “Nonparametric search,” *Journal of Economic Dynamics and Control* 17, 771-84, 1993.
- [8] Morris H. DeGroot: “Some problems of optimal stopping,” *Journal of the Royal Statistical Society. Series B (Methodological)* 30, 108-22, 1968.
- [9] Morris H. DeGroot: “Optimal statistical decisions,” McGraw-Hill, New York, 1970.
- [10] Babur I. De los Santos: “Consumer search on the Internet,” mimeo, 2008.
- [11] Glenn W. Harrison and Peter Morgan: “Search intensity in experiments,” *Economic Journal* 100, 478-86, 1990.
- [12] Gerald Häubl, Benedict G. C. Dellaert, and Bas Donkers: “Tunnel vision: local behavioral influences on consumer decisions in product search,” forthcoming in *Marketing Science*.
- [13] Han Hong and Matthew Shum: “Using price distributions to estimate search costs,” *RAND Journal of Economics* 37, 257-75, 2006.

- [14] Carl A. Kogut: “Consumer search behavior and sunk costs,” *Journal of Economic Behavior and Organization* 14, 381-92, 1990.
- [15] Richard Manning and Peter Morgan: “Search and consumer theory,” *Review of Economic Studies* 49, 203-16, 1982.
- [16] John J. McCall: “Economics of information and job search,” *Quarterly Journal of Economics* 84, 113-26, 1970.
- [17] José Luis Moraga-González and Matthijs R. Wildenbeest: “Maximum likelihood estimation of search costs,” *European Economic Review* 52, 820-48, 2008.
- [18] Peter Morgan and Richard Manning: “Optimal search,” *Econometrica* 53, 923-55, 1985.
- [19] Dale T. Mortensen: “Job search, the duration of unemployment and the Phillips curve,” *American Economic Review* , 84762, 1970.
- [20] Jennifer F. Reinganum: “A simple model of equilibrium price dispersion,” *Journal of Political Economy* 87, 851-58, 1979.
- [21] Donald B. Rosenfield and Roy D. Shapiro: “Optimal adaptive price search,” *Journal of Economic Theory* 25, 1-20, 1981.
- [22] Michael Rothschild: “Searching for the lowest price when the distribution of prices is unknown,” *Journal of Political Economy* 82, 689-711, 1974.
- [23] Andrew Schotter and Yale M. Braunstein: “Economic search: an experimental study,” *Economic Inquiry* 19, 1-25, 1981.
- [24] Joep Sonnemans: “Strategies of search,” *Journal of Economic Behavior and Organization* 35, 309-32, 1998.
- [25] Martin L. Weitzman: “Optimal search for the best alternative,” *Econometrica* 47, 641-54, 1979.
- [26] 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.