This article was downloaded by: [143.89.88.156] On: 19 October 2016, At: 07:19
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Marketing Science

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

Zooming In on Choice: How Do Consumers Search for Cameras Online?

Bart J. Bronnenberg, Jun B. Kim, Carl F. Mela

To cite this article:
Bart J. Bronnenberg, Jun B. Kim, Carl F. Mela (2016) Zooming In on Choice: How Do Consumers Search for Cameras Online?. Marketing Science 35(5):693-712. http://dx.doi.org/10.1287/mksc.2016.0977

Full terms and conditions of use: http://pubsonline.informs.org/page/terms-and-conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2016, INFORMS

Please scroll down for article—it is on subsequent pages

INFORMS is the largest professional society in the world for professionals in the fields of operations research, management science, and analytics.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

MARKETING SCIENCE
Vol. 35, No. 5, September–October 2016, pp. 693–712
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

http://dx.doi.org/10.1287/mksc.2016.0977
© 2016 INFORMS

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

Zooming In on Choice: How Do Consumers Search for Cameras Online?

Bart J. Bronnenberg
Tilburg University, 5037 AB Tilburg, Netherlands, bart.bronnenberg@uvt.nl

Jun B. Kim
Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, junkim@ust.hk

Carl F. Mela
Fuqua School of Business, Duke University, Durham, North Carolina 27708, mela@duke.edu

We describe online consumers' search behavior for differentiated durable goods using a data set that captures a detailed level of consumer search and attribute information for digital cameras. Consumers search extensively, engaging in 14 searches on average prior to purchase. Individual level search is confined to a small part of the attribute space. Early search is highly predictive of the characteristics of the camera eventually purchased. Search paths through the attribute space are state dependent and display "lock-in" as the search unfolds. Finally, the first-time discovery of the chosen alternative usually takes place toward the end of the search sequence. We discuss these and other findings in the context of optimal search strategies and discuss the prospects for consumer learning during search.

Data, as supplemental material, are available at http://dx.doi.org/10.1287/mksc.2016.0977.

Keywords: consumer search and choice; online purchase; search dynamics; digital cameras
History: Received: October 16, 2014; accepted: October 2, 2015; Peter Rossi served as the senior editor and Puneet Manchanda served as associate editor for this article. Published online in *Articles in Advance* May 20, 2016.

1. Introduction

Search is often studied with price as the main characteristic of the goods searched (De los Santos et al. 2012, Honka 2014). Considerably less is known about the more complex process of search for multiattribute products. By providing a detailed account of online search behavior for cameras, this paper describes how consumers search for relatively expensive and differentiated durable goods. Specifically, it characterizes the extent of consumer search, the degree to which search is informative of choice, how search changes over time, and which strategies of search consumers use to find decision relevant information.

Although the question of how consumers search online across attributes for durable goods is simple to pose, addressing it using observational data is challenging. First, only a small fraction of consumers search for durable goods at any given point in time, and an even smaller fraction purchase after a possibly lengthy search. So a comprehensive collection of online searches and purchases requires casting a net over a very large number of consumers over an extended period of time. The resulting browsing data captured in this way are potentially enormous, impeding their procurement and processing. Second, URLs browsed, containing the characteristics of the products searched, are typically dynamic or perishable. This requires extracting the information displayed on the pages requested by consumers concurrently with their search and choice activity.

To address these challenges, past research has aggregated or filtered the data. Search data are often aggregated to the domain level (e.g., visit incidence to a specific retailer), obscuring search within those domains (Huang et al. 2009).[1] In other instances, search data are confined to a particular domain (Kim et al. 2010), abstracting from search across multiple sellers.

These aggregations can lead to partial observation of search. For instance, browsing data confined to a single retailer results in observations of shorter search sequences if the consumer also searches across other retailers, price comparison engines, review sites, and manufacturer websites. Per consequence, the analyst infers high search costs to rationalize the limited search activity.[2]

[1] Baye et al. (2013) find that a significant part of search occurs within a given site.

[2] Whereas it is challenging to observe full search in the field, it is possible to observe full search in a controlled environment such

RIGHTSLINK

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

To alleviate aggregation and filtering, this paper uses a unique consumer-centric panel data set of online search for cameras, constructed to record all consumer browsing activities across- and within-web domains. In particular, we augment and integrate the consumer browsing data with product attribute data collected concurrently from the identical Web pages the consumer browsed, thereby yielding a more complete characterization of consumer search. Using these search data, we find the following.

First, online search for cameras, as defined by the sequence of domain-item Web pages visited, is extensive across products, sellers, and time. Prior to buying a digital camera, a consumer conducts an average of 14 online searches, navigating over three brands, six models, and four domains during six sessions spread over two weeks.

Second, even with extensive search activity, search is confined to a surprisingly small region of the attribute space and the camera chosen almost always lies in the subspace spanned by cameras searched but not chosen. This is true even during early search, and more so during late search. Furthermore, it is common for consumers to revisit alternatives that have previously been searched. Nearly a third of all searches are revisits and such behavior is strongest for alternatives ultimately chosen. About 70% of consumers search the camera they purchase online at only one online retailer. It is thus not common for consumers to first choose a specific camera model and then "price shop" this camera across retailers. By contrast, it is very common for a consumer to search different cameras at a given retailer.

Third, within the small subspace searched, a consumer's path through the attribute space displays strong state dependence. The difference between attribute levels searched and eventually chosen decreases substantially during search. Specifically, consumers "zoom in" on attribute levels from both above and below toward choice. At the same time, the mean attribute level searched remains close to constant throughout the search process. Other dynamics are evident during search. The chosen alternative is typically "discovered," i.e., first searched late in the search. Next, consumers tend to use more generic keywords during earlier search phase and shift to specific keywords later in the search phase. This evidence is also suggestive of discovery of choice late in the search process. Finally, the calendar time lapsed between pages searched accelerates as choice approaches, i.e., searching the first camera takes more time than searching the last camera in a typical sequence.

Our finding that the average attribute levels searched are highly informative of those chosen implies that search is useful in identifying preference heterogeneity. The conclusion that preferences revealed during search are highly similar to those revealed by choice also seems to support empirical research that formalizes search and choice under a unified utility framework (Chen and Yao 2014, Koulayev 2014, Honka and Chintagunta 2016, Kim et al. 2016).

With respect to the nature of the consumer search process, our findings imply that consumers search for more attributes than price and that they do not appear to use a fixed-sample optimal search strategy. Under fixed-sample search, the search sequence within the search set is immaterial to consumers. Therefore, one expects to see absence of state dependence among the searched options and a uniform distribution of the discovery of choice. By contrast, we find state dependence in the path through the attribute space and late discovery of choice, both of which are consistent with sequential search (see, e.g., Kim et al. 2016).

Our empirical findings also bear implications for learning about preferences or alternatives. Because consumers search only a small part of the attribute space, and final attributes chosen are almost identical to the mean level of attribute searched through the entire process, evidence for updating of preferences during search seems limited.

Finally, and substantively, our findings suggest that conditioning on early search allows marketers to better anticipate purchases and make recommendations to yield better outcomes for consumers and firms. As a prime candidate for such intervention, it might be possible to exploit heterogeneity in preferences (identifiable from early search) to improve targeting and recommendations of high value cameras.

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature on empirical research on consumer search. Our data are discussed in §3. Section 4 contains the main body of our study on consumer search behavior. Section 5 concludes with a discussion and suggestions for future work.

2. Literature

2.1. Descriptive Characterizations of Search

Several descriptive studies of online consumer search use domain-level comScore panel data (Johnson et al. 2004, Park and Fader 2004, Huang et al. 2009, De los Santos et al. 2012). Using homogeneous goods

as the lab experiment. The advantages of using laboratory data to characterize consumer search have been argued by Brown et al. (2011). Yet online search can differ considerably from search in the laboratory, because the set of goods searched is endogenous by nature and is not set by the experimenter. Moreover, transactions are real and costly in observational settings. In part for the same reason, recent research using eye tracking in supermarket settings has emerged (Gidlöf et al. 2013). However, search over long periods of time or across retailers remains difficult to replicate with in-store eye tracking.

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

such as specific book titles and CDs, these papers report that consumer search is relatively infrequent and typically limited to one or a few online sellers. For instance, Johnson et al. (2004) find that the frequency of site visits is in the range of one or two per month.

Huang et al. (2009) consider several product categories including cameras. They observe that consumers visit between two and five retailer domains while requesting 60–254 pages across all domains. The extent of within-domain search is unclear because consumers are likely to search for multiple products or categories at merchandisers such as Amazon.com and Walmart.com, i.e., they might be searching other categories concurrently.

2.2. Primitives of Search

Another strand of literature assesses the trade-off between the cost of acquiring additional information against the expected incremental benefit arising from finding a better alternative. This literature builds on decision theoretic models developed in information economics, in particular Stigler (1961) and McCall (1970). For instance, in the case of sequential search, consumers condition on the best option searched and decide whether to sample another (Kim et al. 2010), whereas in the case of fixed sample search, consumers precommit to sampling a fixed set of alternatives (De los Santos et al. 2012, Honka 2014). To organize our discussion, we focus on the nature of goods searched, the type of consumer search strategy, and search dynamics.

Single-Attribute Search. A large part of the recent literature on search is in the context of similar goods, with consumers searching a single attribute (such as price) across sellers, assuming either a process of sequential search (Brynjolfsson et al. 2010, Wildenbeest et al. 2014), or of fixed sample search (Moraga-Gonzalez and Wildenbeest 2008, De los Santos 2012, Lin and Wildenbeest 2013, Moraga-Gonzalez et al. 2013, Honka 2014), or contrasting the two (Hong and Shum 2006, De los Santos et al. 2012, Honka and Chintagunta 2016).

Overall, search for these similar items is found to be relatively short. For example, Hortaçsu and Syverson (2004) infer consumers typically search for just one mutual fund company before purchasing from the category even though the fee dispersion coefficient across companies is about 67%; about 75% of consumers search only one store for books (De los Santos et al. 2012); and a typical consumer visits an average of two–three online retailers for MP3 players (De los Santos et al. 2016). Estimates of search costs are commensurately high. For example, De los Santos et al. (2016) infer the median search cost to be nearly $27.86 for consumers to search for an additional page of results.

Recent empirical studies (De los Santos et al. 2012, Honka and Chintagunta 2016) suggest that consumer price search behavior for similar goods such as books and insurance across sellers is more consistent with a fixed sample search strategy. We complement this literature with more detailed data and for more highly differentiated goods.

Multiattribute Search. For most types of products, consumer search involves more than searching for the lowest price across retailers. It typically also involves searching for the preferred set of attributes (e.g., Hortaçsu and Syverson 2004, Kim et al. 2010, Seiler 2013, Koulayev 2014). Yet even in the context of search for differentiated products with multiple attributes, the literature typically suggests search is short. For example, Moraga-Gonzalez et al. (2013) find that consumers typically visit fewer than 3 sellers for memory chips. More generally, when search engines return multiple pages, consumers typically do not venture beyond one or two pages (Brynjolfsson et al. 2010, Chen and Yao 2014, Koulayev 2014). Huang et al. (2009) document that consumers typically visit 3.2 domains in the camera category. Moraga-Gonzalez et al. (2015) study purchases of cars and note that the median consumer does 2 dealer searches. As above, the short search length implies high search costs. For example, Ghose et al. (2012) estimate that the search cost associated with an additional page of hotels on a travel site is about $40.

A number of papers in this research stream have also integrated choice and search data to improve inference regarding the valuation of goods (Seiler 2013, Koulayev 2014, Kim et al. 2016). Koulayev (2014), for example, shows that preferences estimated from a static discrete choice model that ignores search decisions is biased because choice sets are formed endogenously. That is, those consumers who sort by price in search are more price sensitive in choice. By observing more complete search strings, as we do here, it is possible to more fully characterize the consistency between attributes searched over a session and the attributes chosen and draw an inference on the relationship between search and choice.

Search Dynamics. Empirical research on the temporal characteristics of search is sparse, perhaps because of the partial observation of consumer search used in the past. Notable exceptions are De los Santos et al. (2016) and Koulayev (2014) who explain that consumers revisit items that were searched previously because of learning or nonstationary search costs. Koulayev (2014) shows that decreasing reservation values, in the presence of increasing search cost, can also explain a consumers' tendency to purchase previously visited items under sequential search, thereby generalizing and offering a possible explanation to the finding in De los Santos et al. (2012).

RIGHTSLINK

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

Several studies focus on consumer learning about the distribution of attributes (Bikhchandani and Sharma 1996, Adam 2001). However, although these papers represent important theoretical advances, empirical insights on search dynamics remain limited.

3. Data

3.1. Sources

In this section, we provide an overview of our data sources and collection approach. We collect search data for digital cameras guided by the assumption that the returns to search are high in this category because cameras are differentiated durable goods with ample innovation in features between purchases.

The data are comprised of three main sources. First, we collect comScore log files that contain complete URL-level, browsing histories, and a separate file for all online transactions for a set of online panelists between October and December of 2010.[3] Observation of complete URL-level data allows us to infer the set of online retailers that consumers visit and the set of cameras they browse within and across retailers during their online search process.

Second, we download camera product pages from the three largest online retailers (Amazon, Best Buy, and Walmart) between mid-October and December 2010. This file captures the attributes of cameras including the price offered at those three stores. We impute the nonprice characteristics of a camera at other retailers using the same camera at the three largest online retailers.

Third, we collect daily price histories of select cameras from price tracking websites to impute missing prices in the second data source if any as well as prices at other retailers.

3.2. Collection Approach

Because the entire comScore log file (even over a three month observation period) is many terabytes, it is simply not feasible to sample all 2,000,000 households in the global comScore panel, or even the 250,000 households in the U.S. comScore panel, and query their logs for camera search at all possible URL addresses. Hence, in close collaboration with comScore, we used the following approach to collect our data.

1. *Construct a list of panelists looking for a camera.* We flag all comScore panelists who browsed a minimum of one camera at one of the three largest U.S. online retailers: Amazon.com, Walmart.com, and Bestbuy.com. Over the data collection window of October–December 2010, this yields approximately 12,700 panelists.[4]

2. *Collect camera attributes.* During the same data collection window, we collect the daily product detail pages for all digital cameras for sale at the top three U.S. online retailers.

3. *Collect all browsing behavior.* After the data collection window ended, comScore supplied the *entire* URL-level browsing history in the three preceding months for the identified panelists. Importantly, this includes browsing on all domains, including all other online retailers, manufacturer's sites, camera review and magazine sites, price engines, and so forth. The three largest retailers jointly capture more than 75% of online camera transactions in our sample.

3.3. Integration

We merge the comScore data with product attribute data to obtain a data set containing each searched camera in terms of Web domain, time, and product attributes. Although we relegate details of how these data are processed and merged to the appendix, it is worth noting that the process involves a number of innovations, including the parsing of vast amounts of URL titles, language processing, and text filtering. The process yields a sequence of retailer cameras visited by consumers on which we base our analysis of search.

The integrated data thus prepared (1) are consumer centric, (2) cover all domains, (3) with complete, within-domain URL history, and (4) match product attribute data to each URL. This enables us to observe the complete browsing behaviors of consumers across domains and within domains, subject to the important caveat that it is possible for consumers to augment online search with off-line search.

3.4. Final Sample Selection

Because our analysis characterizes how search and choice are related, we focus on consumers who both search for cameras online and purchase at least one camera at any online outlet. Over the three months, we capture search on online purchases by 967 consumers.[5]

[3] comScore obtains transaction data (purchases and prices) from the panelist's browser at the time of transaction.

[4] This means our data do not cover searches and purchases for users that browse exclusively outside the largest three retailers. However, our data do include purchases and searches at smaller retailers conditional on a user browsing at least once at the largest three retailers (see step 3 of our collection approach).

[5] Given the data collection approach, almost all of our search behavior is conducted by U.S. panelists. Of the 2 million comScore panelists, about 250,000 are comScore U.S. panelists. We observe 977 purchases among the 967 panelists. This means that 0.4% of the U.S. panelists purchased a camera over the 2.5 months that we observe. This implies an annual online purchase rate of 1.9% for digital cameras. For comparison, the Consumer Electronic Association reported that 15% of consumers intended to purchase a digital camera in 2012 (Tillmann and Ely 2012). Forrester reports that the online market share of nongrocery retail sales in 2009 was 11% (http://www.statista.com/statistics/203043/online-share-of-total-us-retail-revenue-projection/). The resulting camera purchase rates online is roughly estimated at 1.6%, which is close to the 1.9% in the comScore data.

RIGHTSLINK

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

Table 1 Market Shares of Top Selling Brands

Brand	Units	Dollars	Dollar share
Canon	158	68,873	0.329
Nikon	187	49,852	0.238
Kodak	236	22,023	0.105
Sony	96	20,963	0.100
Fuji	93	13,526	0.065

Notes. Total number of units in sample $N = 967$. Units and dollars are in sample. The list of brand names is not exhaustive to conserve space.

Table 1 lists the choice shares of the top brands in our data. In dollar terms, Canon is the largest selling brand with Nikon second and Kodak third. The fourth largest selling brand in our data is Sony. These brands are also the largest selling brands in 2010 for the U.S. market (Tarr 2011).

3.5. Unique Searches

We consider unique product searches. Each time a panelist explores another option, as defined by a domain-item combination, we count a search. If consumers refresh the current page, no new search step is counted. If, on the other hand, they visit another site prior to revisiting a specific domain item, we count both the visit and the revisit as separate steps in the search.

In addition, we focus on product pages and therefore exclude specific search queries that direct a consumer to those pages as a search step (e.g., a Google search for a particular model). Moreover, we exclude multi-item pages (because it is hard to impute which item is searched on a multi-item page). As a result, our ensuing insights may be a conservative characterization of the degree of search.

3.6. Attributes

For our empirical analysis, we define three groups of attributes. The first set includes product attributes, specifically SLR (single lens reflex), image stabilization, face detection, movie capability, display size, pixel count, price, and zoom. The second set includes dummy variables for whether a given camera is one of the three largest selling brands: Canon, Nikon, and Kodak. Searching a different brand means that the three attribute dummies have the value 0. The third set includes dummy variables for whether a given camera was searched at one of the three largest online sellers: Amazon, Best Buy, or Walmart. Searching at a different domain means that the three dummies are 0.

4. Online Search Behavior

4.1. How Extensive Is Online Search?

Figure 1 portrays the distribution of individual searches for cameras conditional on purchase. In each plot, the consumer distribution of the extent of search shows a long right tail, which is truncated at the 95th percentile of the distribution. The extent of search is measured in terms of six variables. The first four variables cover the number of unique domains, brands, models, and products (defined as any combination of domain, brand, and model). The fifth variable counts the number of searches conducted by a consumer, whereby a search is defined as a combination of a product URL and time. The number of search steps is larger than the number of unique products searched, because consumers can revisit products. The final variable is the elapsed time between starting a search and buying a camera. In principle, elapsed time is left censored by the data collection method, which considers the three months prior to the October–December 2010 window in which the data were collected. However, in the online appendix (available as supplemental material at http://dx.doi.org/10.1287/mksc.2016.0977) we show that, in the vast majority of cases, search does not last beyond a month. Hence, censoring is likely not very frequent in the data.

Search across domains is not uncommon, spanning 3.5 online domains such as Amazon.com or BestBuy.com. Nonetheless, 41% of panelists are loyal to one domain. On average, 72% of a household's search volume is concentrated within the household's most visited domain.

Search across brands and models is also common. At the product level, a household searches an average of 2.8 brands and 6.4 models prior to purchase. Only 39% of consumers search one brand, and only 21% search one product. Thus, for many consumers, the final brand purchased is not known prior to search. At the brand-model-domain-time level, the typical household search sequence contains 13.9 searches (of which 9.0 are unique domain-brand-model combinations). The difference between total and unique item searches is due to consumers revisiting the same product domain.

The duration between initiating and concluding search conditional on purchase averages just over 15 days and spans an average of 5.9 sessions.[6] Only 25% of consumers make a purchase in one session. Hence, just as search extends over many products in actual decision making about camera purchases, search also extends over time.

In sum, search is extensive in each dimension studied with a high degree of heterogeneity across consumers, potentially suggesting a large variation in search costs and/or consumer uncertainty.

4.2. Search and Choice

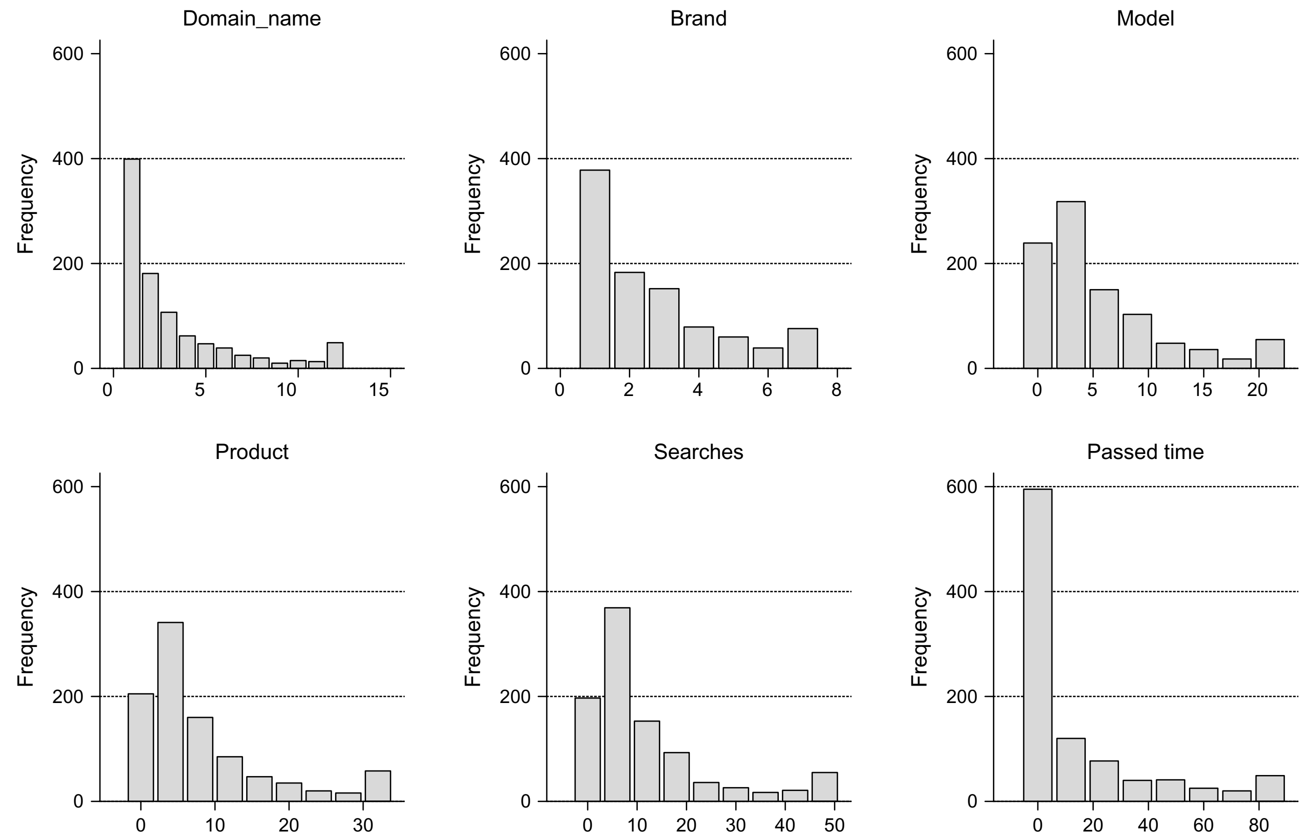
We assess the link between search and choice by leveraging our attribute level information and focusing on attribute levels searched versus the attribute levels

[6] We define a session as a 30 minute interval spent on the machine that records search (Ulmer 2010).

RIGHTSLINK

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

Figure 1 **The Extent of Search**



Notes. The sample distribution of the extent of online search across households shows a large right tail. The small spikes at the right tail are due to truncation and collapse of the right tail above the 95th percentile. Searches are counted as the number of camera-model-domain transitions for each machine. Passed time is calendar time between first search and purchase counted in days.

chosen. To generate more insights into this relationship, we first link search to choice, second link revisiting options in search to choice, and third consider the fraction of the joint attribute space that is actually searched by any given individual prior to choice.

The Link Between Attribute Levels Searched and Chosen. To formalize, let camera j searched by consumer i have attribute k. In our context, we collect data on 14 camera attributes (see §3.6), including retail domain. Denote x_{ijk} as the attribute level of the camera searched, and y_{ik} as the attribute level of the chosen camera. Define $\bar{x}_{ik}$ as the mean attribute level searched by i. This consumer-specific mean is taken across all searches carried out by i, including revisits.[7]

[7] Choice is not counted as a search, although the chosen alternative may be part of a search stream. To fix definitions, if the search sequence is $j = 2, 5, 8, 2, 9$ and choice is 5, then the average $\bar{x}_{ik}$ is computed over $j = 2, 5, 8, 2, 9$ and not $j = 2, 5, 8, 2, 9, 5$. We present several robustness analyses below. When we refer to search without the chosen alternative, we computed averages with $k = 2, 8, 2, 9$, dropping 5 altogether. When we refer to first time search only, we compute averages across $j = 2, 5, 8, 9$, dropping the second visit to option 2. The resulting outcomes of our analysis are highly similar. See the online appendix.

For each continuous camera attribute, Figure 2 shows a scatter plot of the observed values of $\bar{x}_{ik}$ and y_{ik}, along with the best fitting line for these observations and the 45° line for reference. From the closeness of the best fitting line and the 45° line, the average level searched for an attribute is close to an unbiased predictor of the level chosen for each of the four attributes. In addition, the fit of the observations to a straight line is quite good. For instance, the R^2 for the (log of the) price attribute is 0.626 and the R^2 for the other continuous attributes ranges from 0.327 to 0.649. Thus, the average price level searched is highly predictive of the price point at which the consumer buys. Similar observations are made for pixel count, size of display, and the zoom magnification. The online appendix shows that the figure is robust to eliminating revisits to the same camera from the search sequence.

The previous analysis can be generalized to dichotomous attributes via a regression framework that associates attributes searched to attributes chosen. Specifically, we estimate the regression equation

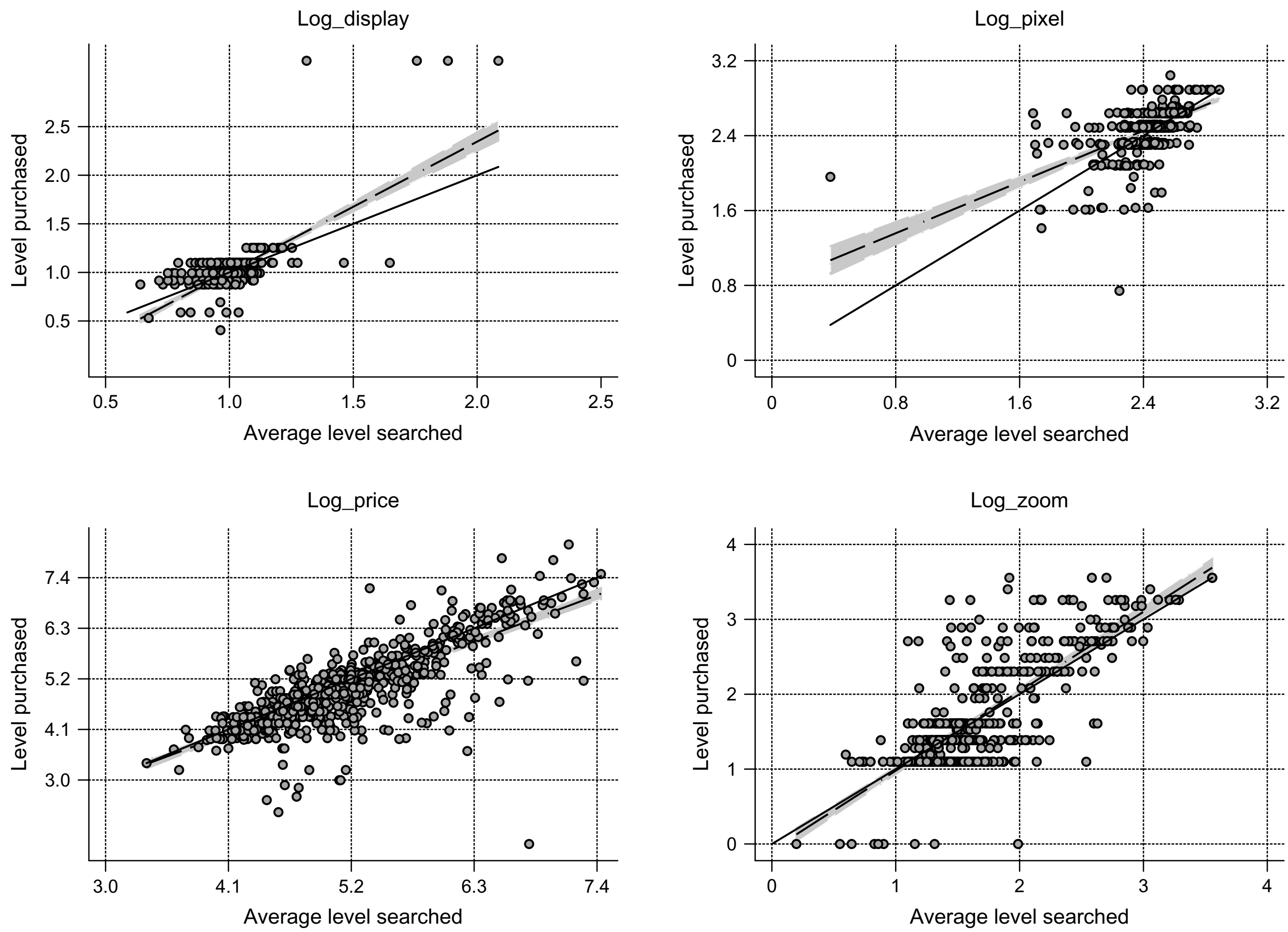
$$y_{ik} = \beta_0 + \beta_1 \bar{x}_{ik} + \varepsilon_{ik}, \tag{1}$$

and test for the joint hypothesis that $\beta_0 = 0$ and $\beta_1 = 1$. When y is dichotomous, Equation (1) is similar to a

RIGHTSLINK

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

Figure 2 **Informativeness of Search on Choice**



Notes. Each scatter symbol represents the within-consumer average of the attribute level searched versus the level chosen by a consumer. The solid line is the 45 degree line. The dashed line is the best linear fit and the shaded area is the 95% confidence interval of the best linear fit. This visualization can only be shown for the continuous attributes pixel, display size, zoom, and price. For dichotomous attributes, regression results are available from Table 2 in the online appendix.

linear probability model. Although we relegate detailed results for all 14 attributes to the online appendix to conserve space, the regression results confirm that the average attribute levels searched $\bar{x}_{ik}$ are highly informative about the attribute level chosen, y_{ik}, for all retail domains and product attributes $k = 1, \ldots, 14$, including dichotomous ones. For example, with respect to choosing a Nikon camera, the regression coefficient (standard error) of searching for a Nikon camera is $\beta_1 = 1.001$ (0.032) and the intercept is $\beta_0 = -0.019$ (0.012). This means that if a consumer searches 10 products, of which 3 are Nikon, then the odds of a Nikon being chosen are also about 3 in 10 (1.001×0.3). Likewise, the regression coefficients for Amazon.com are $\beta_1 = 1.027$ (0.034) and $\beta_0 = 0.069$ (0.015). For most attributes, the estimates of β_0 and β_1 are close to $\beta_0 = 0$ and $\beta_1 = 1$, although the joint hypothesis of H_0: ($\beta_0 = 0, \beta_1 = 1$) is frequently rejected by the data owing primarily to the high power of the test.

Variation in Searched Levels Within and Across Consumers. We next consider the span of attributes searched by a consumer relative to the span of attributes searched across consumers and present in the data. Denote x_{ijkd} as the attribute level of the kth attribute of camera j searched by i in the dth stage of the search sequence. We use the concept of a search stage to normalize the length of search activity across consumers. In particular, the search stage d for each consumer is measured in deciles by dividing each string of searched cameras in 10 equal parts. For example, if an individual searches 30 cameras, each search stage consists of three cameras; the first three searched in the first decile, the next three searched in the second decile, etc. In the general case, if $t = 1, \ldots, N_i - 1$ counts the searches made by individual i with the choice at $t = N_i$, the unbiased normalization of consumer i's search into deciles $d(t, N_i)$ is defined as

$$d(t, N_i) = \text{ceil}\left(\frac{10 \times (t - r(0,1))}{N_i - 1}\right), \tag{2}$$

for $t = 1, \ldots, N_i - 1$, where $r(0, 1)$ is a random uniform number on $0-1$, and ceil is the operator that rounds up

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

to the next whole number.[8] We subsequently use this concept and definition of search decile d throughout the paper.

The variation in searched levels for each attribute k is decomposed by running the following regression:

$$x_{ijkd} = \beta_{ik} + \gamma_{kd} + \delta_{ikd} + \varepsilon_{ijkd}. \tag{3}$$

This equation decomposes attribute levels searched into an individual fixed effect β_{ik}, a common trend on the search decile, γ_{kd}, and the interaction of individual and search sequence effects δ_{ikd}.

Estimating this regression for each attribute, we find that across-consumer variation in the attribute levels searched is uniformly high. For example, 71% of the variance in the logarithm of prices searched is due to individual fixed effects $\beta_{i,\text{price}}$, i.e., is across-consumer variation. Although the total variation in prices searched is large across consumers and search deciles, the vast majority of this variation is across consumers. Across attributes, the median fraction of across-consumer variation is 40%. We provide further details in the online appendix.

We next report the typical range of attribute levels searched by an individual relative to the total empirical observation range across individuals, truncated at the 1st and 99th percentile for robustness. These ranges are only defined for the continuous attributes: display size, pixel count, price, and number of magnifications in zoom. Because we expect to observe the largest extent of exploration of the attribute space for consumers who search most, we conservatively restrict our analysis to consumers whose search sequences involve at least 5 or more cameras ($N = 592$). Table 2 indicates the median consumer searches only a very small fraction of the full range of prices (0.109), and zoom magnifications (0.241). Similar observations hold for the other two attributes. If one considers just the 4 continuous attributes (out of 14 total) as spanning an orthogonal attribute space, then the region that is covered by the modal consumer's camera search is of a volume of 0.2% of the total attribute space. Thus, consumers search only a tiny fraction of the attribute space spanned by the orthogonal combination of the attribute levels that are supplied.

[8] This definition randomly assigns searches at the border between two consecutive deciles, properly randomizing short search strings in a balanced way. For instance, assume a consumer who searches three cameras and buys the fourth one. According to Equation (2), the first search is assigned to decile ceil$(10 \times (1 - r(0,1))/3)$, which has a 0.3 probability of being 1, 2, or 3, and 0.1 probability of being 4. The second search is randomly assigned to decile ceil$(10 \times (2 - r(0,1))/3)$, which has a 0.2 probability of being 4, a 0.3 probability of being 5 or 6, and a 0.2 probability of being 7. Finally, the last search is randomly assigned to ceil$(10 \times (3 - r(0,1))/3)$, which has a 0.1 probability of being 7, and 0.3 probability of being 8, 9, or 10. Thus, by using this definition, each decile has a chance of 0.3 ($= 3/10$) of being used while at the same time ensuring that early searches are in lower deciles and later searches are in higher deciles.

Table 2 Attribute Level Ranges Searched Across Households

	Household percentiles				
	10	25	50	75	90
Display	0.059	0.176	0.294	0.647	0.882
Pixel	0.137	0.151	0.302	0.576	0.842
Price	0.022	0.042	0.109	0.358	0.805
Zoom	0.021	0.069	0.241	0.517	0.797

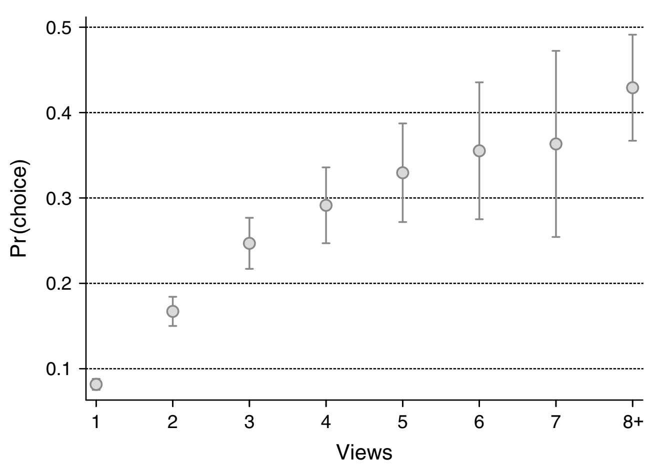
Notes. We report the distribution across consumers of the attribute range of searched items relative to the robust (1%–99%) empirical observation range of that attribute in the data. For example, for the pixel attribute the 25th percentile has the value 0.162. This means that 25% of individuals search less than 16.2% of the empirical observation range of prices.

Because the typical search region of the attribute space is small, we conclude that consumer search behavior is chiefly characterized by the inspection of similar options. Across consumers, however, search explores a wide variety in attribute space.

Repeat Search Behavior. Consumers frequently revisit previously searched items. Return visits account for 31% of all search activity in the data. Revisiting a product more than 2 times accounts for 17% of total search volume across consumers, and revisiting more than 3 times accounts for about 11%. These revisit rates suggest that there is a reason to revisit options once searched, i.e., that the relevant information about a product is not necessarily acquired or remembered in full on the first visit.

We consider whether revisits are informative about choice. Figure 3 represents the coefficients of a linear probability regression of the choice dummy y_{ij} on the number of prechoice visits n_{ij} to each unique alternative. The regression accounts for consumer fixed effects to account for differences in search set size.

Figure 3 The Probability of Choice as a Function of Viewing Multiple Times



Notes. The intervals reflect the 95% confidence bounds of the prediction that an item is chosen given that it has been viewed $n = [1, \ldots, 7, 8^+]$ times. The graph accounts for consumer fixed effects.

RIGHTSLINK

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

The empirical relation between revisiting a particular camera multiple times and choosing it is very strong, statistically and substantively. Observing a consumer searching a camera 4 times, instead of once, increases the choice probability from 0.082 to 0.291. The choice probability increases monotonically in the number of repeated searches. Thus, consumers tend to ultimately purchase cameras they view frequently.

4.3. Evolution of Search

In the preceding section, we documented that online search is predictive of choice and that it takes place in a small region of the attribute space. We now focus on the patterns of search within the small region of the attribute space that consumers explore as they progress through search.

State Dependence in Searched Levels. We begin by testing for state dependence in the search path through the attribute space. More practically, for each of the 14 attributes defined in this study, we estimate the following dynamic panel regression equation:

$$x_{ikd} = \alpha_{ik} + \rho_k x_{ikd-1} + \varepsilon_{ikd}, \tag{4}$$

where x_{ikd} is the average level of attribute k in the dth search decile of consumer i's search sequence. This regression cannot be estimated through ordinary least squares methods or fixed effects regression methods because of the correlation between the estimates of the individual level fixed effects α_{ik} and state dependence parameter ρ (see Nickell 1981). We use the Arellano-Bond Generalized Method of Moments (GMM) estimator (Arellano and Bond 1991) with the finite sample correction of the standard errors suggested by Windmeijer (2005).[9]

Table 3 outlines the results.[10] State dependence in the attribute levels of searched options is strong. For instance, the probability of searching a Nikon rises by 0.250 if that brand was also searched during the previous decile compared to searching another brand. The associated t-statistic is 10.0. Of the top selling domains, searching a camera on Walmart.com is estimated to have a carryover of $\rho = 0.461$ ($t = 17.7$). The associated t-statistics for the estimates of ρ_k are all greater than 2. These results are highly robust and similar levels of state dependence are observed when using searches, rather than search deciles, and when excluding multiple visits to the same option, i.e., using first-time visits only. In each case and for each attribute, we reject the null hypothesis of no state dependence.

[9] For details, see Roodman (2009). It is noted that the least squares estimates with fixed effects are generally close to the reported Arellano-Bond estimates.

[10] For brevity, Tables 3 and 4 report only the results for the largest retailers and domains.

Table 3 State Dependence in Attribute Search

	Intercept	First lag	χ^2	p	N
Canon	0.156 (0.006)	0.236 (0.024)	97.474	0.000	560
Nikon	0.160 (0.007)	0.250 (0.025)	100.425	0.000	560
Kodak	0.103 (0.005)	0.282 (0.024)	140.157	0.000	560
SLR	0.107 (0.005)	0.294 (0.025)	136.817	0.000	552
Image stabilization	0.470 (0.015)	0.115 (0.026)	20.217	0.000	552
Face detection	0.789 (0.025)	0.134 (0.028)	23.008	0.000	552
Movie	0.826 (0.025)	0.114 (0.027)	17.415	0.000	552
Log_display	0.776 (0.026)	0.247 (0.026)	92.601	0.000	552
Log_pixel	1.810 (0.071)	0.266 (0.029)	86.250	0.000	552
Log_price	4.071 (0.156)	0.224 (0.029)	57.874	0.000	549
Log_zoom	1.200 (0.043)	0.262 (0.026)	98.688	0.000	551
Amazon.com	0.197 (0.007)	0.227 (0.025)	81.073	0.000	560
BestBuy.com	0.066 (0.004)	0.448 (0.025)	319.119	0.000	560
Walmart.com	0.098 (0.006)	0.461 (0.026)	325.703	0.000	560

Notes. Standard errors are in parentheses. The χ^2 statistic belongs to the Wald test of overall model fit. The number of observations covers the number of searches with observations in at least two consecutive search deciles. That is, if searches of a given household occurred in deciles 3, 5, and 9, then that household is not represented in this table. If they occurred in deciles 3, 4, and 9 (this could happen purely by chance given the definition of a search decile), then the same household will be represented in this table. We report results for the largest brands and domains, but our analysis involves all of them. For example, if the brand dummy for Kodak equals 0, the cameras searched by that consumer and decile are of any brand not equal to Kodak, and not just one of the other brands in the table, i.e., a Canon or a Nikon. The online appendix reports on robustness of this table to using searches rather than deciles.

Convergence Toward Choice. We next consider convergence in search. That is, we test whether searched attribute levels become increasingly similar to the chosen attribute level as search progresses. We do so in two ways, both intended to highlight different features of the data.

First, we test whether attribute levels searched late in the process are more informative of chosen attribute levels than those considered early in search. To this end, we employ a variation of the regression in Equation (1) where the average attribute level $\bar{x}_{ik}$ for household i and attribute k, is replaced by a weighted average

$$\bar{x}_{ik} = \sum_{d=1}^{10} w_{kd} x_{ikd}, \tag{5}$$

RIGHTSLINK

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

Table 4 Decay of Effect of Attributes Searched Early versus Late on Choice

	Intercept	Slope	Carryover	σ	$\chi^2(2)$	p	N
Canon	0.002 (0.006)	0.896 (0.017)	0.193 (0.047)	0.172	43.943	0.000	914
Nikon	0.006 (0.006)	0.966 (0.016)	0.193 (0.036)	0.168	4.547	0.103	914
Kodak	0.019 (0.007)	0.946 (0.016)	0.106 (0.045)	0.179	13.148	0.001	914
SLR	−0.003 (0.004)	0.963 (0.014)	0.349 (0.037)	0.106	10.540	0.005	880
Image stabilization	0.040 (0.014)	0.930 (0.020)	0.219 (0.046)	0.247	12.564	0.002	880
Face detection	0.101 (0.021)	0.904 (0.021)	0.373 (0.031)	0.120	26.302	0.000	880
Movie	0.130 (0.021)	0.875 (0.021)	0.489 (0.026)	0.095	43.776	0.000	880
Log_display	0.014 (0.019)	0.990 (0.018)	0.501 (0.022)	0.080	1.269	0.530	879
Log_pixel	0.057 (0.055)	0.979 (0.022)	0.260 (0.055)	0.116	2.936	0.230	880
Log_price	0.333 (0.103)	0.916 (0.020)	0.484 (0.067)	0.434	53.112	0.000	913
Log_zoom	0.031 (0.026)	0.981 (0.016)	0.307 (0.038)	0.263	1.508	0.471	875
Amazon.com	0.038 (0.010)	0.934 (0.021)	0.335 (0.048)	0.252	14.415	0.001	914
BestBuy.com	0.005 (0.005)	0.917 (0.015)	0.049 (0.035)	0.149	30.105	0.000	914
Walmart.com	0.027 (0.008)	0.956 (0.014)	0.119 (0.044)	0.185	14.532	0.001	914

Notes. Maximum likelihood estimates assuming a normal distribution of the regression residuals. Standard errors are in parentheses. The column $\chi^2(2)$ is $2(LL_1 - LL_0)$, where LL_0 is the log of the likelihood with the intercept = 0 and slope = 1. Note that the carryover coefficients are well below 1, implying that later search is more informative about choice than earlier search. The number of observations covers the number of searches with at least one search prior to choice and at most one purchase per household. We report results for the largest brands and domains, but our analysis involves all of them. For example, if the brand dummy for Kodak equals 0, the cameras searched by that consumer and decile are of any brand not equal to Kodak, and not just one of the other brands in the table, i.e., a Canon or a Nikon.

with x_{ikd} representing the average attribute level for machine i on attribute k in search decile d. The $\bar{x}_{ik}$ are recency weighted averages by choosing the weights w_{kd} as in Bronnenberg et al. (2012)

$$w_{kd} = \frac{(\beta_{2k})^{11-d}}{\sum_{n=1}^{10}(\beta_{2k})^n}. \qquad (6)$$

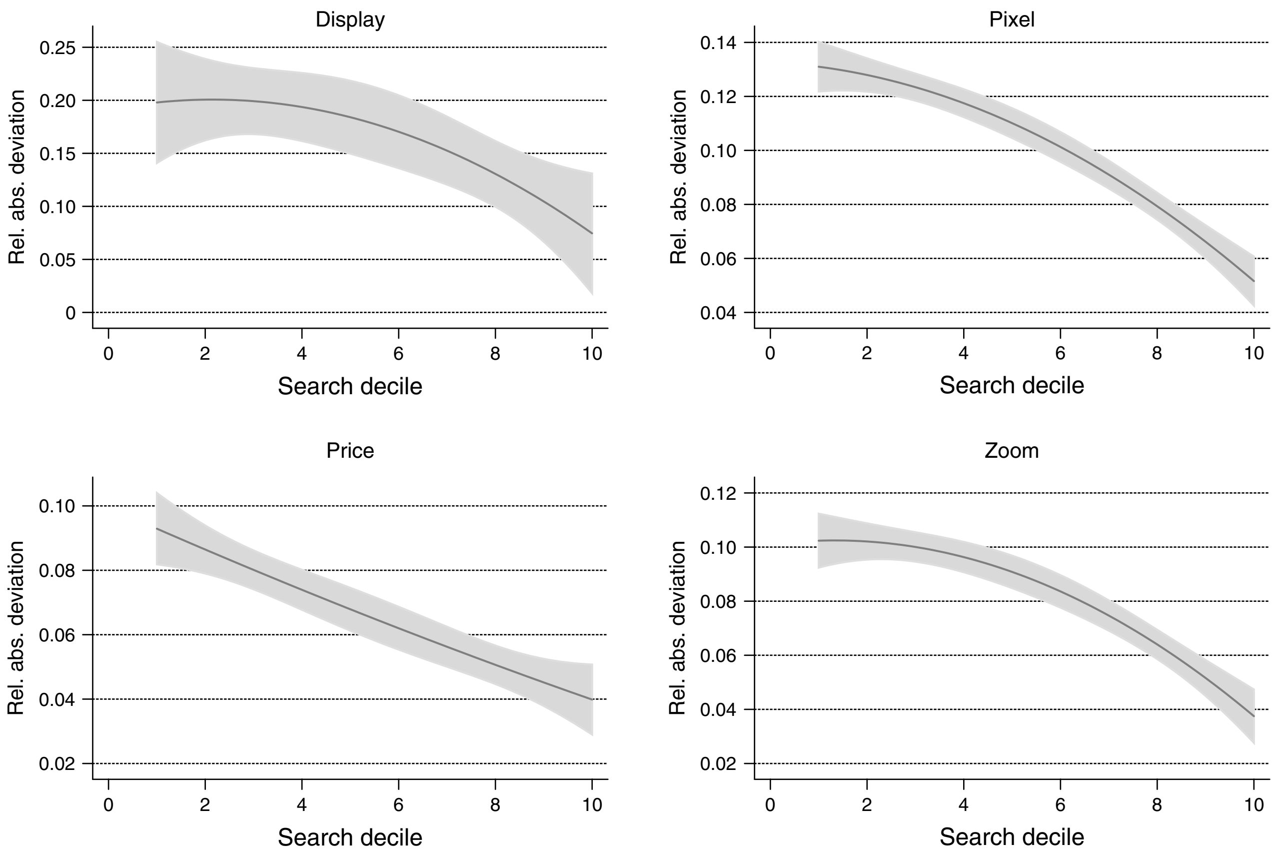
The parameter β_{2k} measures dependence on the past. Specifically, when β_{2k} is close to 0, only the last searched attribute levels are informative of choice and not those searched early. By contrast, when β_{2k} is close to 1, the attribute levels searched during each decile are all equally informative of choice, and the weights in each decile are 1/10. More generally, if the carryover coefficient $\beta_{2k} < 1$ then late search is more informative of choice than early search.

Table 4 reports that typical values for the carryover coefficient β_{2k} are in the range 0.05–0.50 across different attributes k. Placing this finding in perspective, a value of 0.25 near the middle of this range implies that the decile closest to choice has a weight of 0.750 and the fifth decile has a weight of 0.003. For each attribute, β_{2k} is significantly less than 1. In other words, late search is most predictive of choice. In the online appendix, we suppress revisits as a robustness check, and find that the range of values for β_{2k} is 0.41–0.73. This is consistent with the earlier claim that revisits are informative of choice and, by extension, of the chosen levels of the attributes. After eliminating revisits, late search is still more informative of choice than early search, because all β_{2k} are significantly and substantively smaller than 1. For instance, a value of 0.55 near the center of the above range implies that the most recent decile has a weight of 0.45 and the fifth decile has a weight of 0.04. Table 4 also shows the outcome of the likelihood ratio test with the null that $\beta_0 = 0$ and $\beta_1 = 1$. As before, this hypothesis is frequently rejected because of the power of the test, although for each attribute the estimates for

RIGHTSLINK

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

Figure 4 **Convergence to Chosen Attribute Levels**



Notes. The horizontal axis represents search deciles, and the vertical axis shows a measure of the range of attribute levels searched by a given individual relative to a robust measure of the attribute level variation in the data. Using the notation in the paper, if y_{ik} is the attribute level of choice and x_{ijkd} is the attribute level of product j that was searched in decile d, then, for each $[i, k, j, d]$, we first construct the vector $\delta_{ijkd} = |y_{ik} - x_{ijkd}|$, and take averages across j to obtain δ_{ikd}. We also construct the 98% empirical observation range in searched levels for each attribute from the 1st and 99th percentile of the search data in the attribute k. Denote this range by R_k. It is supposed to capture the robust empirical observation range for the attribute k. Then, the vertical axis of the graphs is δ_{ikd}/R_k. The graph thus represents the relative deviation from chosen attribute levels as a function of completion of search. The best fitting quadratic polynomial plus the 95% confidence interval is shown.

β_0 and β_1 are close to the hypothesized values under the null.

Second, we consider the evolution of the absolute deviation between searched and chosen attribute levels over search deciles. We compute for each consumer i, attribute k, and decile d, the mean absolute difference between attributes searched and attributes chosen

$$\delta_{ikd} = \frac{\sum_{j \in d_i} |x_{ijkd} - y_{ik}|}{N_{id}}, \tag{7}$$

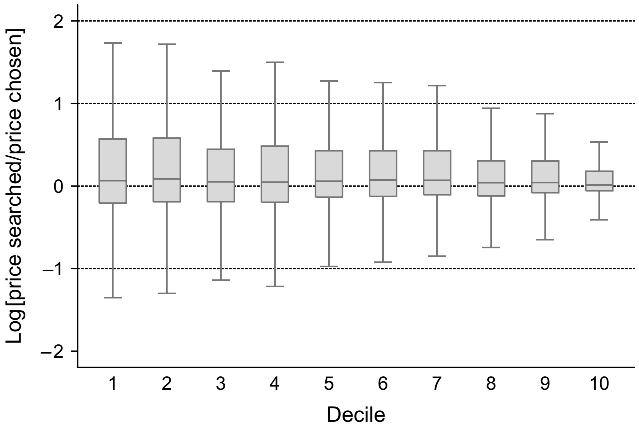
where N_{id} is the observed number of searches by i in search decile d (i.e., the total number of searches for consumer i divided by 10 up to the integer constraint), and d_i is the set of cameras searched during i's decile d. Figure 4 depicts the average δ_{ikd} across consumers, i.e., δ_{kd} for selected k across all d. We normalize these mean absolute deviations by the range of attributes searched by consumers, truncated at the 1st and 99th percentile for robustness.

We find that the difference between searched and chosen attribute values diminishes considerably in later stages of search. In particular, for the continuous attributes such as display, pixel, price, and zoom, the mean absolute deviation in the first decile of search is two to three times higher than in the final decile. For example, at the start of search, the prices of searched cameras deviate from those of chosen cameras on average by just 8% of the full price range. Near the end of search, this reduces further to 4%. The same insight is obtained when we disregard revisits or drop the purchased camera from the prechoice search sequence. The online appendix further shows that the figure is also robust to subsampling consumers who search across multiple domains. Thus, the funnel does not seem to be primarily generated by a single seller recommending ever more similar cameras.

The online appendix offers evidence that switching domains is not accompanied by a widening of the funnel. Controlling for machine and decile fixed effects, we find that switching domains has a small negative effect on the width of the funnel (the δ_{ikd}) for display, pixel, price, and zoom. Rather than leading consumers

RIGHTSLINK

Figure 5 **Convergence to Chosen Attribute Levels**



Notes. This graph represents the distribution of the log of relative prices (the ratio of price searched and price of the item chosen) by decile. The average of the log of relative price is constant and close to 0, but the variance in prices searched drops continuously during the search. This rendition of the graph uses all observations including revisits and is robust to excluding revisits.

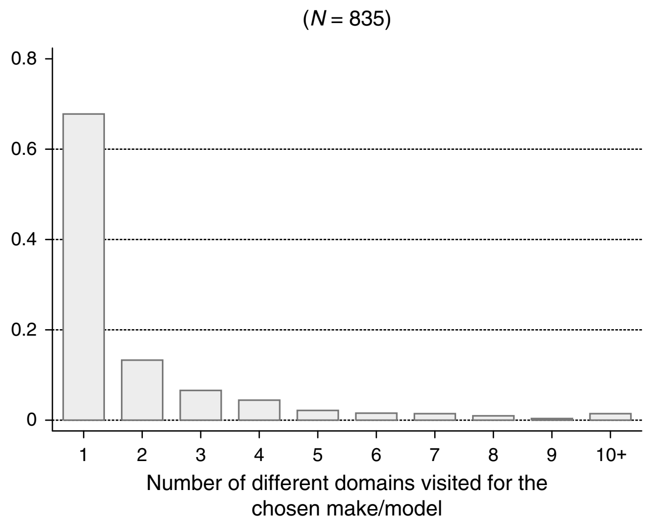
away from their choice, a domain switch either has no effect or a small homing effect toward the chosen option.

We can enrich the description of this convergence by looking at the empirical distribution of searched attribute levels by decile. Figure 5 shows the distribution of prices searched relative to prices of the chosen camera by decile. Although the average relative price searched is close to 1 in each decile, indicating the average price searched early in search reflects the price of the ultimate choice, the variance of the relative price searched decreases throughout the search.

One simple explanation for the observed attribute convergence to the chosen levels is that the later stages of search are dominated by consumers' price shopping their preferred alternatives across different retailers. To explore further, Figure 6 portrays the distribution of the number of different domains that the consumer visits and searches the ultimately chosen brand-model combination. Because the large majority of consumers search the chosen brand-model at only 1 retailer, the convergence we see is not mainly driven by price search conditional on identifying the brand and model of choice but by consumers searching increasingly similar cameras. The online appendix shows that the price dispersion for the same brand model of camera across retailers in this industry is low, suggesting that consumers may have little incentive to price shop across retailers upon finding a brand model of interest.

When Is the Chosen Item First Discovered? Figure 7 reports the distribution of first discovery of the chosen camera across search deciles. Because we are interested in first discovery, we eliminate revisits. The graph shows the distribution of first discovery deciles for

Figure 6 **The Distribution of the Number of Retailers Queried for the Chosen Camera**



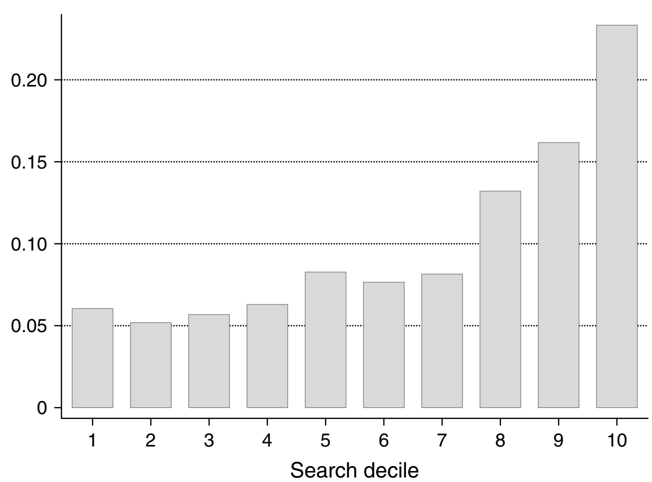
Notes. This graph shows the distribution of the number of different domains that were visited for the chosen brand-model combination. Approximately 70% of consumers search the chosen camera at only one seller. Slightly more than 10% of consumers search the chosen camera at two sellers, etc.

$N = 810$ consumers. Although some consumers find their purchased camera early on, most consumers buy a camera that was first searched late in the search sequence, i.e., in the high search deciles.

Given the late discovery of the chosen camera in search sequence, many consumers stop search immediately or soon after they have identified the product they like.

Are Search Queries the Same Throughout Search? In addition to capturing Web page visitations through URLs, our data capture consumer search queries.

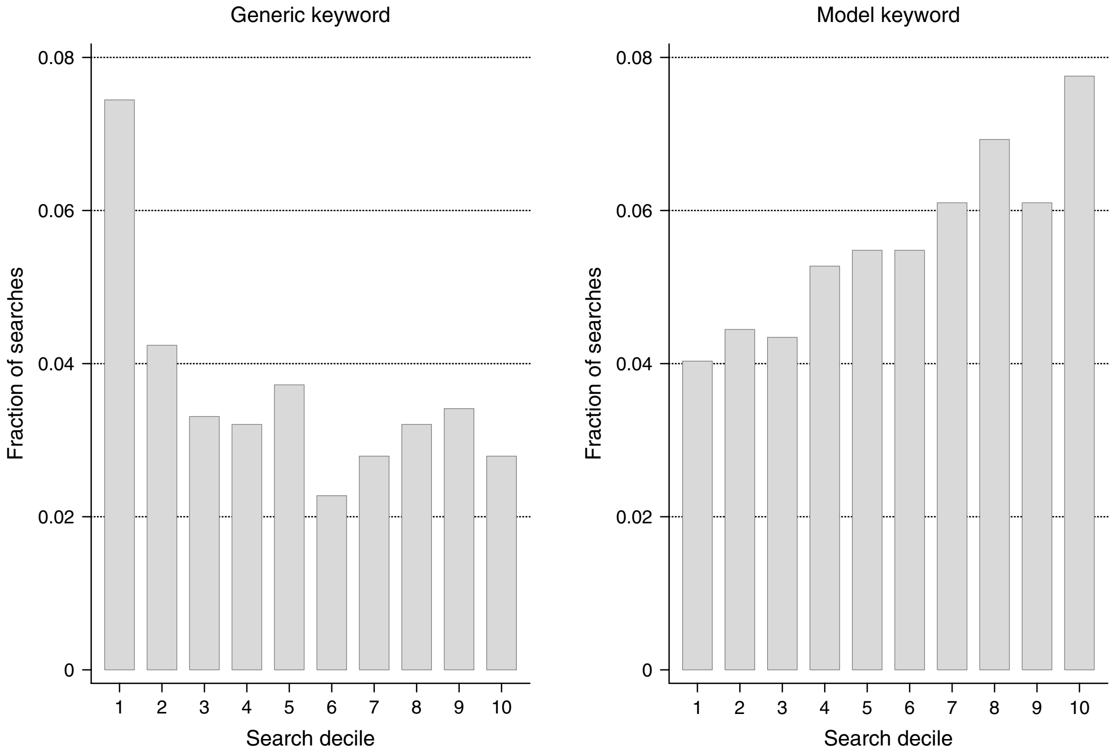
Figure 7 **Likelihood of Discovery of the Chosen Option by Decile**



Notes. This graph represents the distribution of the timing of discovery of the chosen product across search deciles. Product is defined as a unique combination of domain, brand, and model. The deciles are based on first visits.

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

Figure 8 **Keyword Search by Search Decile**



Note. This graph represents the fraction of search sequences for which a keyword search was requested by the consumer by decile and keyword type.

A search query is a keyword query consumers enter in a search engine (e.g., www.Google.com) or retailers (e.g., www.Amazon.com). We analyze the distribution of search query keywords over search deciles. To this end, we first classified camera-related keywords into four categories: generic (e.g., digital camera), brand (e.g., Nikon), product line (e.g., Coolpix), and model (e.g., P520).

Figure 8 suggests that generic keywords are used more often in early stages of search. In 7% of search sequences, the consumer uses a query with a generic keyword to aid navigation during the first stages of search. This drops to 3% in later stages of search. Queries with a model-related keyword are used more prominently later in the search process.

Thus, in sum, the total incidence of keyword queries does not decrease over search stages. However the composition of keywords changes: as search proceeds, consumers use less generic but more concrete keywords. This pattern is also consistent with late discovery of the camera ultimately purchased. The online appendix shows that visits of nonretail websites (such as general information, manufacturer, or price comparison websites) constitutes less than 20% of search activity and drops slightly in late search. The large majority of search for digital cameras is carried out at retailer websites.

Search Time Intervals. Finally, we consider the time elapsed between the first search in each decile in hours. The online appendix reports that among consumers who conduct at least 10 searches, the amount of elapsed time per decile drops substantially from slightly under 90 hours on average to complete the first decile of searches, to about 25 hours to complete the final decile of search. Interpreting the elapsed time as an indirect measure of time consumers spend acquiring information about the products, this result may be consistent with a learning-by-doing hypothesis, or that more prototypical comparisons take less time to process (Sujan 1985).

5. Discussion

5.1. Search and Choice

Our findings are summarized in Table 5. In this section, we synthesize our findings and discuss their implications regarding search and choice, optimal search strategies, and learning.

Our analysis of the comScore panelists' search behaviors yields several insights regarding search and choice. First, consumer search appears informative about consumers' heterogeneous preferences for the following two observations. The searched attribute levels correlate highly with those of the item ultimately chosen (item 2). In addition, whereas the attribute levels searched by

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

RIGHTSLINK

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

Table 5 Summary of Findings

	Search and choice
1.	Search is extensive in the number of options searched and time commitment involved.
2.	Attribute search is centered around the attribute levels ultimately chosen.
3.	A third of search volume stems from revisiting previously searched alternatives.
4.	Revisits are more likely for chosen alternatives.
5.	Individual search spans a small part of the attribute space searched in the aggregate.
	Evolution of search
6.	There is considerable state dependence in attribute levels searched.
7.	The range of attribute levels searched narrows as search proceeds.
8.	The chosen product is discovered late.
9.	As search proceeds, generic queries decrease and item ones increase.
10.	Search accelerates toward its conclusion.

a given consumer are typically narrowly centered around what is chosen (item 5), collectively consumers explore a diverse set of cameras. Consumer search data therefore provide rich information about consumer heterogeneity. The strong correspondence between search and choice supports recent work that links search and choice in demand models (Ghose et al. 2012, Chen and Yao 2014, Honka and Chintagunta 2016, Kim et al. 2016). Moreover, the mean level of attributes searched remains constant and is equal to the level of choice throughout search. This constitutes support for studies that assume that the utility function during search and at the time of choice are the same. Whereas the consideration set literature has historically assumed separate utilities for consideration and choice (Bronnenberg and Vanhonacker 1996, Moe 2006), our findings suggest that consumers search and choice decisions are based on the same utility function.

Second, consumers consistently search a small region of the attribute space. On one hand this suggests consumers enter the search process with well-established preferences. An alternative possibility is that sellers are highly effective at suggesting preferred cameras to consumers based on their limited search histories and at affecting consumer search decisions. However, the majority of consumers search cameras across multiple retailers who would find it impossible to anticipate consumer preferences prior to the incidence of first search. In the online appendix, we show that the funnel in Figure 4 is robust to conditioning on searches across multiple retailers where retailer learning about consumer preferences is likely limited.

Third, consumers frequently revisit products they searched earlier (item 3). In the context of consumer "recall," De los Santos et al. (2016) suggest consumers can recall because of learning; as the option value of learning decreases, so too does the value of search. Another explanation is increased search costs, which further rationalize consumers' decisions to choose previously visited alternatives (Koulayev 2014). A cognitive explanation for the revisits may be incomplete attention or high search costs of concurrently processing all attributes relative to sequential processing; implying that one visit to an alternative does not resolve all of its uncertainty.

5.2. Optimal Search Strategies

Inferences about search and choice primitives strongly depend on assumptions about which search strategy the consumer uses (see, e.g., Honka and Chintagunta 2016). Recent studies (De los Santos et al. 2012, Honka and Chintagunta 2016) conclude that consumers adopt a fixed sample search strategy during price search for products such as specific books and car insurance policies. In this section, we address the same question in the context of search for multiattribute goods such as digital cameras.

Under a fixed sample search strategy, the consumer first establishes a search set of a given size and precommits to searching it entirely. Therefore, one would expect the search sequence to be random within the set. However, we find (item 6 in Table 5) that the order in which the consumer moves through the attribute space is nonrandom. By contrast, state dependence and a search funnel offer support for the existence of a sequential search process. Sequential search theory predicts that it is optimal for the consumer to search ("select") products in decreasing order of reservation utility, which indexes the attractiveness of searching an option (see, e.g., Honka and Chintagunta 2016, Kim et al. 2016).

More specifically, using the approach in Kim et al. (2010), the attractiveness of search or reservation utility can be written as $z = V + \sigma\zeta(c/\sigma)$, where V is the expected utility, σ is the standard deviation of the unknowns in the indirect utility function, c is the cost of search, and ζ is a monotonically decreasing function.[11] Assuming constant σ and c, an ordering on reservation utility z is equivalent to an ordering on V. In addition, conditional on consumer preference (or, utility coefficients or β) V is a function of observed product attributes. Therefore, for any given consumer, ordering on V promotes ordering of attribute tuples in a linear utility framework. If the consumer searches in order of z, and thus V, then this in turn produces state dependence in attribute levels searched, as we observe. Therefore, there is a direct relation between optimal sequential search and state dependence in the attribute space.

Second, another finding that supports sequential search strategy is the discovery timing of the chosen option in the search sequence (item 8). A consumer using a fixed sample search strategy is indifferent to the

[11] In this conceptualization, product utility is defined as $u = V + e$ with $\mathrm{var}(e) = \sigma^2$.

RIGHTSLINK

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

order of search within the set and therefore we expect to observe a uniform distribution for the discovery timing. By contrast, sequential search promotes late discovery of the chosen option, with the exact location of the discovery dependent on the realization of the utility draws. For instance, when the consumer searches options that are similar with respect to observable characteristics (item 5), the expected utilities V do not vary greatly across options searched. In this case, we expect the highest utility alternative to be associated with a large positive utility shock. In addition, such shocks are also likely to terminate search. Thus, we expect the choice and search termination to be closely located and hence the choice to be discovered late. This is exactly what is found in Figure 7. If we test for the null of uniform discovery timing of the chosen camera, this hypothesis is rejected at any significance level ($\chi^2(9) = 251$, $p = 0.000$).

We conclude that state dependence, lock in, and discovery timing of choice collectively reject fixed sample search process. By contrast, the reported patterns are consistent with a sequential search process.

5.3. Learning

There has been a large theoretical literature on consumer learning and search, dating back to the 1970s (Rothschild 1974). However, empirical work regarding learning in search is sparse. Recent exceptions are De los Santos et al. (2016) and Koulayev (2009) who consider learning in theory-based empirical search models. In the context of consumer search, we make the distinction between learning about preferences and learning about attribute distributions. We discuss both through the lens of our model-free evidence.

First, the objective of learning about preferences is to improve one's private mapping of different attribute levels to utility. For instance, a consumer may look at a number of point-and-shoot cameras to learn that she really wants an SLR. If so, then this consumer will migrate her search from a non-SLR region to a region in the attribute space with many SLRs in later search. Indeed, Adam (2001) discusses the case of learning about preferences in the context of optimal sequential search. After positively (negatively) updating preference on a certain attribute level, the consumer next optimally migrates her efforts to (from) the unsearched items with that attribute level. However, we observe no migration of search (items 2 and 5 in Table 5); for example, Figure 5 indicates that consumers search the same average level of price in each decile. Similar graphs can be drawn for the attributes display, pixel, and zoom. From the absence of migration of search in the attribute space, we conclude that learning about preferences during search does not seem to be dominant. Alternatively, our data seem to suggest that perhaps learning is confined to confirming an existing preference function.

Second, the objective of learning about attribute distributions is to determine, for example, the variance of the attribute distribution (Koulayev 2013). The key pattern in our data that relates to attribute learning is that, compared to late search and choice, consumers initially search options farther away from the chosen option (item 7 in Table 5). This seems consistent with an interpretation that early exploration of the range of attributes is followed by a more focused search later and that some learning about the attribute ranges has taken place. However, an implication of optimal search is that if a consumer believes that the variance of the unobserved utility components is large, all else equal, search sets should be large from the upside potential of each searched item. We can contrast consumers who initially search very diverse items with consumers who search across a more focused set of items and test a relation between search focus and observed search set size. We find that, early exploration of the attribute space, measured as the absolute difference between levels searched in the first decile and choice, does not correlate highly with the size of the search set ($\rho = 0.164$). Additional modeling is warranted to conclude whether consumers engage in learning about attribute distributions. We leave this to future research.

5.4. Managerial Implications and Future Extensions

The set of our findings raises the possibility that marketers can intervene in the search and change its path and outcome. This can be done within and across consumers. Across consumers, early search is still informative about preferences. Hence, similar options can be advertised to those who search for them. Within consumers, one could imagine that incomplete search strings are highly effective targeting or direct marketing variables. For example, during the early stages of the search, advertising might focus on the brand, whereas during the later stages of the search process, it might focus on product lines or models. Later in the search, a retailer can change the order of items presented to more closely align with the preferences exhibited by those who search.

Our analysis can be extended along multiple dimensions. First, we considered only one durable-goods category. Although we expect our results to be informative of search in durable consumer goods categories more generally,[12] search in nondurable categories might

[12] We reason that our key findings will be applicable to product categories that are highly differentiated, infrequently purchased, and relatively expensive. First, past research reports intensive consumer search in highly differentiated (e.g., Schaninger and Sciglimpaglia 1981, Duncan and Olshavsky 1982) or in expensive product categories (e.g., Udell 1966, Newman and Staelin 1972). Second, recent empirical studies (Kim et al. 2010, 2016) report high levels of consumer heterogeneity in the camcorder market, a category close to cameras. Third, behavioral research reports that consumers typically adopt a

RIGHTSLINK

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

be shorter in categories where prices are low and preferences are well formed, such as consumer grocery goods. In the automotive insurance market, Honka (2014) reports that switching costs play a major role in search. Although our data characterize online search, one might expect shorter search in offline environments due to higher search costs (e.g., see Moraga-Gonzalez et al. 2015, who consider the offline search for cars).

Second, still more detailed data could reveal further insights on consumer search. One would expect reviews, in particular, to affect the nature of search because the recommendations of experts and consumers are influential in choice (West and Broniarczyk 1998, Chevalier and Mayzlin 2006). In our data, both review and product detail information exist on a single page, making it impossible to disentangle these effects. Likewise, because multiple product attributes are concurrently presented on a single product page, it is challenging to assess which particular attributes are searched. Hence, we characterize search for multiattribute goods, but cannot observe multiattribute search. Knowledge of attribute search order can be obtained through eye tracking software using computer cameras (e.g., Teixeira et al. 2014).

Third, it is compelling to explore which search patterns are associated with purchase. This analysis might prove useful in characterizing how search relates to purchase incidence and interventions that might impact it.

Last, it is fruitful to develop a structural model of search and choice that fully takes into account not only the observed search sets but also the observed sequence of search in studying consumer search and demand. This additional information should help identify demand primitives better.

6. Conclusions

Procuring and processing a new disaggregate data set covering consumer online search for digital cameras allowed us to obtain a more complete characterization of how search evolves "in the wild." Among the main findings of our paper, we find that consumers are not simply price shopping a preselected camera; (even early) search is highly informative about heterogeneous consumer preferences; search through the attribute space is strongly state dependent; and the chosen option is discovered late.

From a measurement perspective, our findings suggest one can augment choice data with search data to obtain better identification of heterogeneous preferences of consumers. From a theoretical perspective, our findings seem consistent with predictions from sequential search strategy. Finally, from a practitioner perspective, firms can presumably use early search histories to target messages and better recommend alternatives as search proceeds.

heuristic called choice by processing attributes (CPA) during their search process (Russo and Dosher 1975, Jacoby et al. 1976, Bettman and Jacoby 1976). This attribute-driven search process implies dependence among key attribute levels of searched products. Finally, Kim et al. (2010) report results from a different category, camcorders, that are consistent with some of our results. For instance, a typical consumer searches around 11 camcorder options and consumers are inferred to have strongly heterogeneous tastes. Furthermore, consistent with our finding that consumers do not explore the attribute space, camcorders that are viewed together in the same session tend to share the same attribute levels.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mksc.2016.0977.

Acknowledgments

The authors are thankful for comments from Allan Collard-Wexler, Elisabeth Honka, Jean-Pierre Dubé, Pranav Jindal, Ralf van der Lans, Sriram Venkataraman, John Payne, and seminar participants at the Chinese University of Hong Kong, the University of Colorado at Boulder, IDC Herzliya, the Ohio State University, the Marketing Dynamics Conference in 2014, Temple University, the Triannual Invitational Choice Symposium in 2013, and the University of South Australia. The authors are especially grateful to comScore for their generous provision of the data that enabled this research. The first author thanks the Netherlands Foundation for Science (NWO) for financial support [Grant 453-09-004]. The second author acknowledges the support from the Hong Kong Research Grants Council [Project 691613].

Appendix. Data Processing

We discuss the three data sources and our procedures for processing, merging, and preparing the final sample from the three data sets.

A.1. Data Sources

A.1.1. comScore Panel Data. comScore maintains a panel of two million global Internet users who allow the firm to confidentially track and capture their online browsing and transaction behaviors. comScore installs tracking software on its panelists' computers to record the URL of the Web pages rendered at users' browsers. Additional tracking software records panelists' online transactions. In addition, comScore maintains demographic information regarding its panelists. Please visit ir.comScore.com for more details regarding the nature of its panel. These data form the backbone of our analysis. The panel data were collected by comScore between early August 2010 and mid-December 2010.[13]

The comScore log files contain the following information: person ID, machine ID, complete URL of Web page requested, HTML title of the Web pages rendered, domain name, and time of request for the page visited.[14] Person ID is the

[13] Because comScore collects domain level information from its panelists, some work sites prohibit its use. As a result, traffic and search may be undercounted in some instances.

[14] In total, the number of header entries in the file is 23.

RIGHTSLINK

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

unique identifier assigned to a panelist and machine ID is the machine (i.e., computer) specific ID. The entire URL refers to the source address of the Web page rendered at the panelist's browser.[15] Domain name represents the Internet content host identifier. For instance, www.cnet.com and reviews.cnet.com share the same domain, cnet.com. The time of request records the time stamp when a URL was requested.

comScore also provided a data file containing the online transactions made by the panelists between September 1, 2010 and January 31, 2011. Fields in this file include event time, person ID, machine ID, product name, product information, price, and online store domain. The product name is the complete description of the chosen item and typically includes brand name, model number, and other physical characteristics of the product.[16]

Last, comScore afforded us user demographic information. This data set includes the location of computer (i.e., home or work), household-related information such as household income and household size, head-of-household information such as race, employment status, marital status, education level, and location-related information such as country, zip code, and designated market area (DMA).

A.1.2. Retailer Product Page Data. In addition to the data from comScore, we collected daily product information pertaining to the cameras purveyed by three major online retailers of Amazon.com, BestBuy.com, and Walmart.com. To this end, we automate the process of collecting all store-specific unique camera IDs, constructing the URL for each camera Web page, and downloading corresponding product detail Web pages.[17] This product detail page is the actual Web page rendered at the panelists' browsers on a certain date.

We repeat this process on a daily basis from mid-October 2010 until early January 2011. Upon collection of the product Web page, we programmatically parse each file and extract static and dynamic product characteristics. Static camera characteristics are technical and nontechnical attributes that do not change over time such as brand name, zoom, pixel number, and model number. Dynamic attributes are something that can change over time such as prices and consumer reviews.

A.1.3. Additional Pricing Data. Last, we selectively downloaded files from price tracking websites. This online service provides price trajectory of a vast majority of products sold at Amazon.com.[18] This enables us to observe prices for Amazon's camera prices outside our data collection time window.

Table A.1 **Initial Keywords Used to Identify Potential Camera-Related Records in the Data File**

Keyword category	Actual keywords
Generic	camera, dslr, point-shoot, point_shoot
Product line	cyber-shot, cybershot, lumix, coolpix, powershot, finepix, ax250, exilim, easyshare, xacti, stylus, lxus, capilo
Manufacturer	Sony, Canon, Panasonic, Nikon, Samsung, Olympus, Casio, Fuji, Kodak, Leica, Pentax, Polaroid, Ricoh, Vivitar, Konica, Minolta, Yashica

A.2. Data Preparation

By integrating these data sources, we construct a complete set and sequence of camera-related Web pages requested and browsed by a consumer and reconstruct when, where, and what kinds of cameras online consumers browsed before making an online camera purchase. Next, we detail how we extract, transform, and merge the aforementioned three data sources to achieve this aim.

A.2.1. Transaction Data. We first delineate the camera transactions in comScore's online transaction file as this file contains purchases of many goods.[19] To do this, we create a means for filtering noncamera purchases by using the product descriptors in the transaction files. At first, we prepare a set of camera-related keywords as is shown in Table A.1. We consider three different categories of keywords. First, a "generic" keyword is a category-level keyword that directly refers to camera or camera subcategories. Second, a "camera product line" keyword refers to manufacturer's product lines, if any. For instance, one of Sony's camera product lines is "cybershot." Last, our keywords also include names of all camera "manufacturers." We prepare the list of all camera manufacturers from Amazon.com's camera section, which details most camera manufacturers and product line names. Note that keywords in the list are often "duplicative" in the sense that a product name in the transaction file typically contains all types of keywords.[20]

Next, we parse the online transaction file against the prepared camera keyword list to assess whether a record in the transaction file contains at least one camera-related keyword. If a match is found between a description and the keyword, we identify the potential for the purchase record to be a camera transaction. Upon parsing, we manually inspect the product name of each candidate record to assess whether it is indeed a camera transaction. This manual process removes "false positive" records such as "video camera" or "Samsung HDTV." The final number of filtered camera

[15] Parsing URLs is complicated because a given Web page requested by a consumer can contain information from another URL that is not requested by the consumer. For example, advertisements delivered by an advertising network such as doubleclick.net can be placed within a Web page requested by the panelists. These "pushed" URLs are not explicitly requested by the panelists.

[16] For instance, one of the product names in the transaction file is "NIKON COOLPIX S70 12.1MP DIGITAL CAMERA WITH 3.5 INCH OLED TOUCH SCREEN AND 5X WIDE ANGLE OPTICAL VIBRATION REDUCTION (VR) ZOOM (RED)."

[17] Each online store assigns a unique ID for each camera it carries. For instance, Amazon.com assigns a unique ASIN (Amazon standard identification number) to each product at its store.

[18] There are many online websites that track the prices of a vast majority of products offered at major retailers such as Amazon.com and BestBuy.com. We use the price data from camelcamelcamel.com.

[19] The original transaction file contains more than 100,000 transactions across different online outlets during the data collection time period.

[20] For instance, one of the product names in the transaction file is "NIKON COOLPIX S70 12.1MP DIGITAL CAMERA WITH 3.5 INCH OLED TOUCH SCREEN AND 5X WIDE ANGLE OPTICAL VIBRATION REDUCTION (VR) ZOOM (RED)." This product description contains a generic (camera), a product line (Coolpix), and a manufacturer (Nikon) keyword.

RIGHTSLINK

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

transactions is about 977. This filtered set also indicates which machine IDs purchased a camera.[21] The transaction machine IDs serve as "transaction condition" and will be used for processing the browsing file described in §A.2.2.

A.2.2. Browsing Data. In this step, conditional on "transaction condition" in §A.2.1, we extract all camera-related Web pages the panelist requested from the comScore log files. Accordingly, this step is central to determining and reconstructing the browsing path prior to purchase. The major challenge inherent in finding camera search episodes is to identify and infer such records embedded within the very large-scale comScore log files. The log file contains all URLs that record consumer activities across different Web pages and all domains including major retailers (Amazon, Best Buy, and Walmart). To do this, we use the URL or "HTML Title" fields, which are informative about the content on the Web page, enabling us to find camera-related information across an array of domains.[22] In instances where URL fields are generated by Java Server Page (JSP) and not informative of its contents, HTML titles are typically still informative about its contents.[23] In case neither URL nor HTML title is informative about the content of the Web page, we can potentially miss and undercount the camera-related activities.

There are two major steps involved in using these fields to isolate camera-related records in the file. First, we write computer scripts to identify all records that could be potentially camera related. To that end, we identify candidate records by parsing each line in the log files against the aforementioned camera-related keywords in Table A.1. This is done by comparing the URL and HTML title of each record in the log file against our set of keywords. If either the URL or HTML title contains any keyword, we retain those records as potential "camera activity" candidates for the next step of data processing. Conditioned on this step, one must next methodically remove pages that are not related to camera search. That is, we remove "false positive" records from the set of records from the preceding automated step. Such records can occur when (i) URLs from irrelevant websites contain accidental camera-related keywords,[24] (ii) advertising networks generate and push camera-related advertising URLs to user's browsers,[25] and (iii) camera manufacturers offer products from other categories, e.g., "Sony HDTV." We take a two-step approach because it is impossible to isolate all camera search steps manually.[26]

A.2.3. Product Attribute Data. Product data are used to enumerate product characteristics for all searched camera models (e.g., "Nikon Coolpix S70" or "Sony W330"). To do this, we programmatically parse product characteristics such as a manufacturer's brand name, model number, and technical characteristics from the files downloaded from the three online retailers. This generates attributes for roughly 1,400 items across all stores. Specifically, we extract brand name, model number, SLR, zoom, pixel, display size, sensor, video feature, image stabilization, and face detection features.[27]

This operation, at times, yields inconsistent product data (e.g., same or different retailers provided different attribute values for identical models such as different colors). To resolve these inconsistencies, we isolate and remove these conflicts by manually visiting other third-party websites. All told, the resulting product set yielded 750 unique "manufacturer-model" options as well as their attribute values.

A.2.4. Matching Attributes to Browsing. Subsequently, we merge the prepared data in §§A.2.2 and A.2.3 and infer the specific camera models consumers browsed during their online session. Technically, for each record in the browsing data, we check whether any brand-model string we identified

[21] In the data, we observe some panelists with a user ID of "0." The value of "0" may be users who did not post their user IDs during the Internet sessions. Accordingly, the use of user ID may risk undercounting the level of online activities. By contrast, machine ID may confound activities of different panelists who share the machine for camera purchase. This may be a concern if different panelists sharing the computer make different camera purchases during our sample period of three months. Owing to the relatively short time span, we suspect these occurrences are minimal and thus believe machine ID are a more appropriate identifier in our sample and equate a machine ID as one panelist.

[22] For instance, many URL structures across websites are very informative of product category or individual products. Examples are reviews.cnet.com/best-midrange-dslr-cameras/?tag= and www.Amazon.com/Panasonic-DMC-F2K-10-1MP-Digital-Optical/dp/B00395Y9FA?comScorebkw=address. From the above two URLs we can infer that the page behind the first URL is about camera category and the second URL is about a Web page about an individual camera product.

[23] For instance, during our data collection period, BestBuy.com had Web pages generated by JSP. The following is one URL for such an example: http://www.BestBuy.com/site/olstemplatemapper.jsp?id=1218109500127&type=product. Although this URL is not informative of any camera, its corresponding title is, Nikon-Coolpix-12.0-Megapixel-Digital-Camera-Blue-S570.

[24] For instance, we find URLs from some noncamera-related websites contain the keyword "camera", weather.weatherbug.com/common/akwriter.aspx?zip=12540&city=&postal_code=&lang_id=en-us&camera_id=&units.

[25] One example URL by an advertising network is ad.doubleclick.net/adi/sz.us.computers/digitalcameras;tok.... where a URL from doubleclick contains the keyword camera. In addition, some technology-driven online retailers such as Amazon.com also embed their own advertising URL within a requested Web page. An example of Amazon.com's own advertising URL is www.Amazon.com/gp/product-ads/lazyLoad/handler/related-ads.html?ie=UTF8&ASIN=B0035FZJI0, where "B0035FZJI0" is an Amazon SKU for one of the cameras at its online store.

[26] To remove the first and second types of records, we identify all of the unique domain names (over 1,000 in total) and the number of unique visitors to these domains. For each domain with more than 2 unique visitors, we manually visit those Web pages and exclude noncamera-related domains. To remove the third case, we create a list of confounding keywords such as "HDTV," "bag," and "battery." If the URL and HTML title of each candidate record contains any of these, we remove the observation. One concern is the incompleteness of the list of these confounding words. An iterative approach is used to address this concern. After each iteration, we randomly select several hundred records and manually inspect each URL and HTML title. If the manual inspection reveals additional disqualifying keywords, we update our disqualifying keyword list and reprocess the candidate set. We iterate a dozen rounds of random-check/update/extract processes until we no longer identify false positive, noncamera records.

[27] Daily price data preparation is described in detail in §A.2.5.

RIGHTSLINK

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

Table A.2 Example of URL and HTML Title of Records in the Final Browsing Data

Source type	Type	Examples
Retailer	URL	http://www.Amazon.com/Nikon-Coolpix-L22-3-0-Inch-Red-primary/dp/B0034XIL60/r...
	HTML title	Amazon.com%3a-Nikon-Coolpix-L22-12.0MP-Digital
Manufacturer	URL	http://www.nikonusa.com/Find-Your-Nikon/Product/Digital-Camera/26194/COOLPIX-L110.html
	HTML title	N/A
Retailer	URL	http://www.BestBuy.com/site/olspage.jsp?_dyncharset=ISO-8859-1&_dynSessConf=-1616368331...
	HTML title	Canon EOS 550d BestBuy
Review site	URL	http://reviews.cnet.com/digital-cameras/canon-powershot-d10/4505-6501_7-33529078.html
	HTML title	Canon PowerShot D10 Review Digital cameras CNET-Reviews

Note. URL and HTML titles are both abbreviated.

in §A.2.3 is embedded in the URL or HTML title in the browsing record. When the URL or HTML title of a record in the browsing file contains both brand name and model strings we infer that the consumer has browsed such a product during the session. For instance, the first two records in Table A.2 contain the brand and model pair value of Nikon/L22 in the URL and HTML title.

A.2.5. Price Data. Daily prices are predominantly collected by downloading the product detail pages of the three focal online retailers (Amazon.com, BestBuy.com, and Walmart.com) and parsing them for pricing information. After extracting the prices, we match them to the browsing file prepared in §A.2.4. Over our data collection period, we extract over 300,000 prices across product, retailers, and time. However, in some instances the pricing and browsing data do not align for the following reasons.

First, comScore log file data and our product data collection are slightly out of phase; comScore log files commence in September 2010 and product page data collection begin slightly later, in mid-October 2011. In this case, we use an online price history data website for the prices outside our collection time window.

Second, we do not have price information from smaller retailers for whom it would have been impracticable to collect page data, such as Ritzcamera.com. In this case, we impute prices by using the prices of identical products at the three focal retailers from whom we scraped data, predominantly from Amazon.com. That is, we use the Amazon.com data as the major source for imputing prices at other online retailers.

Last, owing to rare events such as poor Internet connections with the retailer's server during our collection, we sometimes are unable to capture prices from the three focal stores within our data collection period. In these cases we impute the missing prices by using the prices of the same product from the adjacent dates at the same retailer (i.e., nearest neighbor imputation).

References

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

Arellano M, Bond S (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Rev. Econom. Stud.* 58(2):277–297.

Baye MR, De los Santos B, Wildenbeest MR (2013) The evolution of product search. *J. Law, Econom. Policy* 9(2):201–305.

Bettman JR, Jacoby J (1976) Patterns of processing in consumer information acquisition. *Adv. Consumer Res.* 3(1):315–320.

Bikhchandani S, Sharma S (1996) Optimal search with learning. *J. Econom. Dynam. Control* 20(1):333–359.

Bronnenberg BJ, Vanhonacker WR (1996) Limited choice sets, local price response and implied measures of price competition. *J. Marketing Res.* 33(2):163–173.

Bronnenberg BJ, Dubé JPH, Gentzkow M (2012) The evolution of brand preferences: Evidence from consumer migration. *Amer. Econom. Rev.* 102(6):2472–2508.

Brown M, Flinn CJ, Schotter A (2011) Real-time search in the laboratory and the market. *Amer. Econom. Rev.* 101(2):948–974.

Brynjolfsson E, Dick AA, Smith MD (2010) A nearly perfect market? *Quant. Marketing Econom.* 8(1):1–33.

Chen Y, Yao S (2014) Sequential search with refinement: Model and application with click-stream data. Working paper, New York University, New York.

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* 43(3):345–354.

De los Santos B (2012) Consumer search on the Internet. Working paper, Indiana University, Bloomington.

De los Santos B, Hortaçsu A, Wildenbeest MR (2012) Testing models of consumer search using data on Web browsing and purchasing behavior. *Amer. Econom. Rev.* 102(6):2955–2980.

De los Santos B, Hortacsu A, Wildenbeest MR (2016) Search with learning for differentiated products: Evidence from e-commerce. *J. Bus. Econom. Statist.* Forthcoming.

Duncan CP, Olshavsky RW (1982) External search: The role of consumer beliefs. *J. Marketing Res.* 19(1):32–43.

Ghose A, Ipeirotis P, Li B (2012) Search less, find more? Examining limited consumer search with social media and product search engines. *Proc. 33rd Internat. Conf. Inform. Systems*.

Gidlöf K, Wallin A, Dewhurst R, Holmqvist K (2013) Using eye tracking to trace a cognitive process: Gaze behaviour during decision making in a natural environment. *J. Eye Movement Res.* 6(1):1–14.

Hong H, Shum M (2006) Using price distributions to estimate search costs. *RAND J. Econom.* 37(2):257–275.

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

Honka E, Chintagunta PK (2016) Simultaneous or sequential? Search strategies in the U.S. auto insurance industry. *Marketing Sci.*, ePub ahead of print August 1, http://dx.doi.org/10.1287/mksc.2016.0995.

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

Huang P, Lurie NH, Mitra S (2009) Searching for experience on the Web: An empirical examination of consumer behavior for search and experience goods. *J. Marketing* 73(2):55–69.

Jacoby J, Chestnut RW, Weigl KC, Fisher W (1976) Pre-purchase information acquisition: Description of a process methodology, research paradigm, and pilot investigation. *Adv. Consumer Res.* 3(1):306–314.

Johnson EJ, Moe WW, Fader PS, Bellman S, Lohse GL (2004) On the depth and dynamics of online search behavior. *Management Sci.* 50(3):299–308.

RIGHTSLINK

Downloaded from informs.org by [143.89.88.156] on 19 October 2016, at 07:19 . For personal use only, all rights reserved.

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

Kim JB, Albuquerque P, Bronnenberg BJ (2016) The probit choice model under sequential search with an application to online retailing. *Management Sci.* Forthcoming.

Koulayev S (2009) Estimating search with learning. Working paper, Columbia University, New York.

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

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

Lin H, Wildenbeest MR (2013) Search and prices in the Medigap insurance market. Working paper, Indiana University, Bloomington.

McCall JJ (1970) Economics of information and job search. *Quart. J. Econom.* 84(1):113–126.

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

Moraga-Gonzalez JL, Wildenbeest MR (2008) Maximum likelihood estimation of search costs. *Eur. Econom. Rev.* 52(5): 820–848.

Moraga-Gonzalez JL, Sandor Z, Wildenbeest MR (2013) Seminonparametric estimation of consumer search costs. *J. Appl. Econometrics* 28(7):1205–1223.

Moraga-Gonzalez JL, Sandor Z, Wildenbeest MR (2015) Consumer search and prices in the automobile market. Discussion Paper, Centre for Economic Policy Research, London.

Newman JW, Staelin R (1972) Prepurchase information seeking for new cars and major household appliances. *J. Marketing Res.* 9(3):249–257.

Nickell S (1981) Biases in dynamic models with fixed effects. *Econometrica* 49(6):1417–1426.

Park YH, Fader PS (2004) Modeling browsing behavior at multiple websites. *Marketing Sci.* 23(3):280–303.

Roodman D (2009) How to do xtabond2: An introduction to difference and system GMM in Stata. *Stata J.* 9(1):86–136.

Rothschild M (1974) Searching for the lowest price when the distribution of prices is unknown. *J. Political Econom.* 82(4):689–711.

Russo JE, Dosher BA (1975) Dimensional evaluation: A heuristic for binary choice. Working paper, University of California, San Diego, San Diego.

Schaninger CM, Sciglimpaglia D (1981) The influence of cognitive personality traits and demographics on consumer information acquisition. *J. Consumer Res.* 8(2):208–216.

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

Stigler GJ (1961) The economics of information. *J. Political Econom.* 69(3):213–225.

Sujan M (1985) Consumer knowledge: Effects on evaluation strategies mediating consumer judgments. *J. Consumer Res.* 12(1):31–46.

Tarr G (2011) Canon, Nikon, Sony top 2010 U.S. DSC market share rankings. Accessed March 20, 2014, http://www.twice.com/digital-imaging/digital-imaging/canon-nikon-sony-top-2010-us-dsc-market-share-rankings/31778.

Teixeira T, Picard R, el Kaliouby R (2014) Why, when, and how much to entertain consumers in advertisements? A web-based facial tracking field study. *Marketing Sci.* 33(6):809–827.

Tillmann K, Ely C (2012) Consumer electronics association digital camera market overview. Consumer Electronic Association, Arlington, VA.

Udell JG (1966) Prepurchase behavior of buyers of small electrical appliances. *J. Marketing* 30(4):50–52.

Ulmer H (2010) Browsing sessions. Accessed October 12, 2014, http://blog.mozilla.org/metrics/2010/12/22/browsing-sessions/.

West PM, Broniarczyk SM (1998) Integrating multiple opinions: The role of aspiration level on consumer response to critic consensus. *J. Consumer Res.* 25(1):38–51.

Wildenbeest M, Giulietti M, Waterson M (2014) Estimation of search frictions in the British electricity market. *J. Indust. Econom.* 62(4):555–590.

Windmeijer F (2005) A finite sample correction for the variance of linear efficient two-step GMM estimators. *J. Econometrics* 126(1):25–51.