Management Science

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

Measuring Benefits from New Products in Markets with Information Frictions

Ilya Morozov

To cite this article:
Ilya Morozov (2023) Measuring Benefits from New Products in Markets with Information Frictions. Management Science

Published online in Articles in Advance 22 Mar 2023

. https://doi.org/10.1287/mnsc.2023.4729

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

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–21
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Measuring Benefits from New Products in Markets with Information Frictions

Ilya Morozov[a]

[a] Marketing Department, Kellogg School of Management, Northwestern University, Evanston, Illinois 60208

Contact: ilya.morozov@kellogg.northwestern.edu, https://orcid.org/0000-0003-1185-4215 (IM)

Received: March 21, 2021
Revised: December 20, 2021; July 24, 2022
Accepted: September 8, 2022
Published Online in Articles in Advance:
March 22, 2023

https://doi.org/10.1287/mnsc.2023.4729

Abstract. I study how much consumers benefit from new products in markets with information frictions. I analyze new products in the U.S. hard drive market, which is characterized by ample product innovation. Using unique click-stream data, I measure the magnitude of two frictions, category consideration and costly search, and show that both play a crucial role in shaping consumer demand. To estimate consumer surplus from new products, I develop a search model that captures both frictions and propose a novel Bayesian estimation method to recover its parameters. I then show that ignoring information frictions leads researchers to underestimate the consumer surplus from new hard drives because it appears that consumers do not value the combinations of attributes these hard drives offer. Partly eliminating frictions, through marketing efforts or market-wide transparency initiatives, can help consumers to more fully internalize the benefits of new product launches.

History: Accepted by Matthew Shum, marketing.

Supplemental Material: The data files and online appendices are available at https://doi.org/10.1287/mnsc.2023.4729.

Keywords: new products • value of innovation • consumer search • Bayesian analysis

1. Introduction

Many industries grow through a constant stream of technological innovations that allow firms to improve existing products or introduce entirely new goods. However, consumers often struggle to stay informed about new product launches, especially in markets with large assortments. To break this information barrier, firms invest millions of dollars in advertising new products and promoting the value these products create.[1] Although industry practitioners see such investments as key to launching new products, the academic literature on this topic has adopted a more simplified view. Much of this literature assumes that consumers become immediately aware of all new products and learn their attributes at no cost (Hausman 1996, Petrin, 2002). Even when consumers do not adopt a new product due to their limited knowledge, this standard approach ascribes new product failures to consumer preferences. By mistaking information for preferences, this approach may obscure the true value of new products for consumers. It may also prevent researchers from studying how firms' marketing efforts can help consumers fully internalize the benefits of new product launches.

This paper has two goals. First, I extend the existing techniques for estimating the value of new goods by accounting for information frictions. I model two specific frictions, *category consideration* and *costly search*, both of which are grounded in empirical evidence. A consumer first chooses which product category to consider within a specific market. By eliminating certain categories, a consumer might not even consider new products regardless of their attributes. In practice, such behavior may arise because a consumer is unaware of certain product types, finds some categories too complex to understand, or chooses to reduce the consideration set to simplify the choice process. Having chosen categories to consider, the consumer learns the basic attributes of all products within these categories but faces residual uncertainty about products' match values. The consumer then examines products one by one to gather missing information, the process I capture using the sequential search model of Weitzman (1979). Because search is costly, the consumer only examines new products if their known attributes seem sufficiently appealing. Importantly, this model allows for the possibility that many consumers do not purchase new products despite the benefits these products offer.

Second, I apply the model to study consumer surplus from new products in the U.S. hard drive market, which is characterized by fast-paced product innovation. I estimate the model using detailed click-stream data from the Comscore Web Behavior Panel, and I measure the extent to which consumers benefited from the introduction of solid state drives (SSDs): a new class of hard drives that substantially increased the read and write speeds of traditional hard disk drives (HDDs). Consumers did not immediately recognize the benefits of SSDs. A survey conducted by Western Digital in 2016, a decade after the commercial introduction of SSDs, revealed that more than 40% of U.S. consumers were still unaware of the SSD technology or could not explain how it differs from HDDs.[2] Therefore, it is possible that some consumers did not even consider the category of SSDs when buying a hard drive. In addition, because of high product turnover and large product assortments, consumers in this market faced a nontrivial search problem. The question I ask in this context is whether and to what extent information frictions affected demand for SSDs, and if so, how that affected the surplus consumers derived from the introduction of SSDs.

The central empirical challenge I need to address is how to separate consumer preferences from information frictions. That consumers do not click on SSDs might be evidence that they have decided not to learn the attributes of these hard drives or chose not to consider the SSD category altogether. It might also reflect that consumers do not like the combinations of attributes that SSDs offer. To measure search frictions, I use a unique click-stream data set from Comscore that contains the universe of web pages visited by consumers in 2016. This URL-level data set is substantially more detailed than domain-level Comscore data used in prior work (De Los Santos et al. 2012, De Los Santos 2018). Importantly, from these data, I know how many and which hard drives each consumer examined before buying a hard drive online. By combining this direct measure of search with data on the actual online purchases, I can empirically separate search costs from consumers' preferences for hard drive attributes (e.g., price, storage capacity, and speed).

A further challenge is how to distinguish consideration from search, as both frictions may prevent consumers from buying SSDs. To separate one from another, I rely on two sets of exclusion restrictions. The first one excludes prices from the category consideration stage but allows consumers to react to price changes when choosing what products to search. That is, a consumer who does not consider SSDs does not know their prices and therefore cannot possibly react when these prices change. Asymmetric reaction to temporary price promotions of SSDs and HDDs then allows me to separate consideration from costly search. The second exclusion restriction builds on novel variables I construct from consumer-level data. For example, I identify consumers who visit informational websites about PC hardware. I then use visits to such websites as a consideration shifter, implying that these consumers might be more likely to consider SSDs due to their expertise in computer hardware. Similarly, I construct a variable capturing how many products consumers search in other product categories (i.e., other than hard drives). Because such a variable partly captures individual differences in the opportunity cost of time, I use it as a search cost shifter. These variables generate additional exclusion restrictions, helping me separate consideration from search.

It is generally difficult to estimate structural choice models with search frictions. The prior work shows that traditional frequentist methods of estimating search models might be slow and numerically unstable (Chung et al. 2019). Estimation becomes even more challenging in my model where I add a consideration stage and allow for rich heterogeneity in preferences, search costs, and consideration probabilities. To overcome this issue, I develop a novel Bayesian estimator that incorporates category consideration and costly search into the standard hierarchical probit choice model (Rossi et al. 2012, p. 75). Compared with frequentist methods, the MCMC sampler I propose is more numerically stable, can more easily handle rich consumer heterogeneity, and scales better to large assortments commonly observed in online markets.

My estimation results show that introducing SSDs raised the surplus of the average consumer by \$3.2, about 3% of the average hard drive price. By contrast, a perfect information model, similar to those used in the prior work on new products, underestimates this surplus change almost by a factor of three. This bias is driven by two distinct effects. On the one hand, the perfect information model inflates the surplus change by incorrectly assuming that consumers know all attributes of SSDs and therefore always know whether some SSDs match their preferences better than HDDs. This is not the case in my model where many consumers do not consider SSDs or remain uninformed about SSDs' attributes. On the other hand, the perfect information model underestimates the surplus change by incorrectly attributing the low market share of SSDs to consumer preferences. The net result of these two opposing effects is that the perfect information model dramatically underestimates the surplus generated by SSDs. Accounting for information frictions, therefore, is crucial for estimating consumer surplus from new products in this market.

Modeling frictions also helps understand how reducing these frictions can affect the surplus generated by new products. My estimates reveal that the magnitude of information frictions substantially affects how much consumers benefit from SSDs. The surplus from SSDs increases by 60% when I partly remove frictions from the estimated model while keeping preferences fixed. In particular, increasing the number of tech-savvy users who frequently consider SSDs raises the surplus from SSDs from \$3.2 to \$4.8. Reducing search costs of all consumers by about 50% further raises this surplus change to $5.0. This observation suggests that one can help consumers benefit from newly introduced hard drives by educating them about SSD technology or adopting a website design that helps consumers discover the SSD subcategory. Put another way, removing the consideration barrier is key: Once consumers consider the SSD category, they will find something they like despite search frictions. I also explore how consumer surplus changes when I reduce consumers' ability to search in a directed way, and I find that the surplus from SSDs reduces substantially. When consumers are forced to search in a random order, they can only discover and learn about SSDs by chance, which reduces the surplus from SSDs to $1.2. This result can be interpreted to mean that currently available search tools help consumers discover and learn about new hard drives, thus increasing consumers' benefits of new product introductions.

One may wonder whether the proposed model is a reasonable way to account for information frictions. I believe that the two frictions I model are both realistic and grounded in empirical evidence. The idea that consumers limit their search to a specific category seems natural: one can imagine a person who is looking for a bottle of wine from a specific region (e.g., French Bordeaux), trying to book a hotel room in a specific neighborhood of a vacation town, or searching for a restaurant of a certain cuisine type (e.g., Italian or Japanese). The first model of such category consideration is proposed by Ching et al. (2009). Manzini and Mariotti (2012) study the theoretical revealed preference properties of such a model. Additionally, consider-then-choose models perform remarkably well at rationalizing anomalies in choice experiments (Manzini and Mariotti 2010), fitting grocery purchase data (Ching et al. 2009), and explaining weak responses to price promotions (Seiler 2013). Similarly, search frictions have been well documented in a large variety of markets (Honka et al. 2019). Given that both frictions seem to be inherent features of consumer behavior, it makes sense to account for both category consideration and costly search when estimating demand for new goods.

This paper contributes to two key strands of literature. The first shares my primary goal of estimating consumer surplus from new goods. Ever since the seminal paper of Hicks (1940), this literature studied the value of new goods in a variety of markets, including those for computers (Bresnahan 1986), breakfast cereals (Hausman 1996), automobiles (Petrin 2002), and books (Brynjolfsson et al. 2003). I show that to obtain plausible surplus estimates, the researcher needs to account for information frictions. The perfect information model misestimates surplus for two reasons: it fails to correctly predict what consumers would have chosen if the new product were not introduced, and it uses an incorrect surplus function that does not account for frictions. Because it is difficult to anticipate the sign of the resulting bias, researchers need to empirically account for information frictions in each application. The model I develop helps researchers account for such frictions, and my application of hard drives illustrates how that can be done in practice. Accounting for frictions also reduces the strong dependence of consumers' choices on unobserved product-specific shocks, thus partly removing the property considered undesirable in the discrete choice literature (Petrin 2002, Ackerberg and Rysman 2005, Berry and Pakes 2007).

The second related strand of literature is on modeling consideration and search. Several authors develop empirical models of consumer search and estimate them using online browsing data (De Los Santos et al. 2012, Honka 2014, Ursu 2018, Donnelly et al. 2022, Moraga-González et al. 2022). My contribution is in extending these models with a category consideration stage, developing a Bayesian estimation method, and using estimation results to study consumer surplus from new products. To the best of my knowledge, the only other paper that jointly models consideration and search is Honka et al. (2017). Their work differs from mine in both the research question and methodology. Although they focus on studying how advertising affects consumer choices, I focus on estimating consumer surplus from new products. They also pursue a different identification argument. To identify whether consumers are unaware of certain brands, which is analogous to category consideration in my model, they directly ask consumers in a survey which brands they are aware of and can recall. By contrast, I examine what consumers do when not shopping for hard drives, and I use this information to construct novel shifters of preferences, category consideration, and search costs. I also rely on an exclusion restriction that removes price from the consideration stage, similar to the identification argument in the consideration model of Ching et al. (2009, 2014). It is worth noting that in addition to category consideration, Ching et al. (2014) model consumer learning about product quality, a channel closely related to search models. However, they do not use their model to measure the welfare benefits of new product launches. In this sense, my paper can be viewed as a complement to their research.

2. Market and Data

2.1. Hard Drive Market

The hard drive market provides an ideal setting to study the value of new products under information frictions. Given product turnover and large product assortments, consumers face a nontrivial search problem. Because repeat purchases are rare in this market, consumers effectively encounter a new search problem each time they return to the market to buy another hard drive. Additionally, this market constantly evolves through technological innovations that allow manufacturers to produce faster and more compact hard drives with greater storage capacities (Christensen 1993, Igami 2017). One recent major innovation came from introducing the conceptually new technology of SSDs. The traditional HDDs consist of circular platters that rotate at high speed, allowing the read-write heads to access information in different platter segments. By contrast, the newly introduced SSDs have no moving mechanical components and are based on NAND flash memory. This ensures SSDs are significantly faster than HDDs, more durable, and more resistant to physical shock.

Despite the promising new technology, consumers have been slow to adopt SSDs. Although online retailers started offering mass-market SSDs in 2010–2011, by 2016, only 22% of hard drive buyers purchased an SSD.[3] One potential explanation is related to consumer preferences. Because SSDs are considerably more expensive, consumers may not be willing to pay for the additional speed that these hard drives offer. Consumers may also prefer to buy a high-capacity HDD rather than a low-capacity SSD if they require additional storage space. Another potential reason for this limited adoption is that consumers are imperfectly informed about SSDs. They may not consider SSDs either because they are unaware of this new technology or because they simply misunderstand it. In fact, a quick online search reveals that PC users often get confused about the SSD technology, claiming that SSDs are incompatible with many modern computers or that SSDs make it impossible to recover data in case of a hard drive failure. Although neither of these claims is true, such misconceptions could prevent consumers from considering SSDs.

2.2. Click-Stream Data

I use click-stream data from the Comscore Web Behavior Panel. The data include the complete browsing histories of 81,418 U.S. Internet users in 2016. These users were chosen at random by Comscore from the sample of 2.5 million U.S. households. Comscore users install software meters on their computers and give Comscore permission to track all their Internet activity. The data set therefore includes the complete browsing history of each user with URL addresses, the history of purchases on major e-commerce websites, and users' demographic variables including age, income, and household size.[4] Importantly, I observe the URL addresses of all visited pages, which makes these data more detailed than most Comscore datasets used in prior work.[5] This highly granular data enable me to recover the exact list of hard drives each user examined and the order in which the user examined them.

To construct the main sample, I identify all users who shopped for hard drives on Amazon.com in 2016. Because more than 75% of hard drive purchases in my data were made on Amazon, this sample captures the majority of searches and purchases in this market. I first recover the complete list of 1,774 hard drives that were available on Amazon in 2016 and identify all users who "searched" at least one hard drive by opening its product page. I define a "search session" as all searches made within a week leading to a purchase; and for users without purchases, I define a search session by finding a week with the largest number of searches. Because users rarely make repeat purchases and conduct all their searches within narrow time intervals, these definitions leave me with a data set where each user has exactly one search session. As a result, I obtain a sample of 2,422 users who searched a total of 4,366 unique hard drives. Only 222 of these users purchased a hard drive from Amazon, implying a conversion rate of 9.2%. Throughout this section, I conduct the analysis using the complete sample. In Section 4, when estimating the structural model, I simplify estimation by focusing on the 100 most frequently purchased hard drives. For additional details of data construction, see Online Appendix A.

Users search relatively little. Despite the large assortment available on Amazon, the average user only searches 1.8 hard drives. This observation suggests that either users face substantial information frictions, or they have strong preferences for certain hard drive attributes that make them reluctant to search beyond a few specific options. Importantly, around 70% of users do not search any SSDs during their search session, and SSDs represent only 21.1% of all purchases and 24.8% of searches in the main data set. In the following sections, I study to what extent this reluctance to search SSDs is driven by preferences, category consideration, or costly search.

2.3. Attributes and Prices

I also collect hard drive attributes from Amazon product pages and daily prices of hard drives from a third-party price tracking website (see Online Appendix A for details). My general strategy, consistent with the identification strategy in Section 3.4, was to collect all attributes that are visible to users on the product category page. I therefore collected information about each hard drive's price, brand, memory type (HDD or SSD), disk type (internal or portable), storage capacity in terabytes, and hard drive speed in megabytes per second. Table 1 summarizes these attributes. SSDs are, on average, faster and more compact than HDDs but offer lower storage capacity and more often require internal installation. Although Seagate and Western Digital offer the largest assortments of HDDs, the category of SSDs is dominated by Samsung and SanDisk. However, the attributes in the two subcategories substantially overlap. Most brands offer both hard drive types, and the assortment includes plenty of fast HDDs and SSDs with high storage capacity. This attribute overlap

Table 1. Attributes of Hard Drives That Were Offered on Amazon.com in 2016

	All drives (N = 1,774)		HDDs (N = 1,346)		SSDs (N = 428)	
Attribute	Mean	Standard error	Mean	Standard error	Mean	Standard error
Storage TB	1.28	1.63	1.51	1.78	0.57	0.65
Speed MB/s	180.1	161.6	94.6	35.8	406.7	145.5
Internal Drive	0.670	0.470	0.606	0.489	0.869	0.338
Brand: Seagate	0.184	0.387	0.230	0.421	0.040	0.196
Brand: WD	0.167	0.373	0.215	0.411	0.019	0.136
Brand: Toshiba	0.067	0.249	0.079	0.269	0.028	0.165
Brand: Samsung	0.058	0.234	0.022	0.145	0.173	0.379
Brand: SanDisk	0.024	0.154	0.001	0.027	0.098	0.298
Brand: Crucial	0.016	0.125	0.000	0.000	0.065	0.248

will prove crucial for identifying category consideration (see Section 3.4).

Table 2 additionally shows that SSDs are almost twice as expensive as HDDs, but both hard drive types go through frequent promotion periods. An average SSD is on promotion 93 of 366 days (7–8 days each month), with an average discount of 9%. HDDs show comparable price variation. As I explain in Section 3.4, these temporary promotions help me identify category consideration.

Users focus their search on hard drives with similar attributes. Table 9 in the Online Appendix illustrates that users consistently search hard drives of the same brand and rarely switch to searching a different brand within the same session. The same observation can be made about other attributes. For instance, 90.4% of HDD searches are immediately followed by another HDD search, and 62.4% of SSD searches are immediately followed by another SSD search. Although this search persistence might reflect that each user only considers one category, it might also reflect that users have heterogeneous preferences for attributes that distinguish HDDs from SSDs (e.g., storage capacity and speed). As I explain in Section 3.4, I leverage this observed search persistence to identify taste heterogeneity.

2.4. User-Level Variables

The Comscore data set gives me a unique ability to study what other websites and pages users browsed outside the category of hard drives. Using these additional browsing data, I construct shifters of preferences, search costs, and category consideration. I briefly describe these shifters here, but interested readers may consult Online Appendix A for details.

2.4.1. Taste Shifters. I construct several variables capturing what other products users searched and purchased apart from hard drives. First, because brand preferences may translate across product categories, brands users searched and purchased in the past might help me predict what brands they will choose when shopping for hard drives. To this end, I collect information on whether users searched or purchased other products from the brands represented in the hard drive category. Second, I identify users who purchased a desktop computer, laptop, video camera, or video game console before shopping for hard drives. Because these purchases might indicate an increased need for storage space, they might predict whether a user will search hard drives with high storage capacity. Finally, I also identify users of file-sharing websites (e.g., torrents) who might need additional storage space for the files they download, as well as the users of cloud storage services (e.g., Google Drive) who might not need much storage space. Using these taste shifters helps me estimate preference heterogeneity, which is especially difficult to do in my application where I do not observe repeat purchases. One could think about these shifters as carrying similar information as panel data, although my estimation uses it in a simplified way, without explicitly modeling multicategory shopping).

Table 2. Depth and Frequency of Price Promotions for SSDs and HDDs

	Price mean	Price standard deviation	Days on promotion	Promotion depth
All drives (N = 1,774)	$140.60	$12.40	77	−9.9%
HDDs (N = 1,346)	$115.30	$7.70	72	−10.3%
SSDs (N = 428)	$220.30	$27.30	93	−9.0%

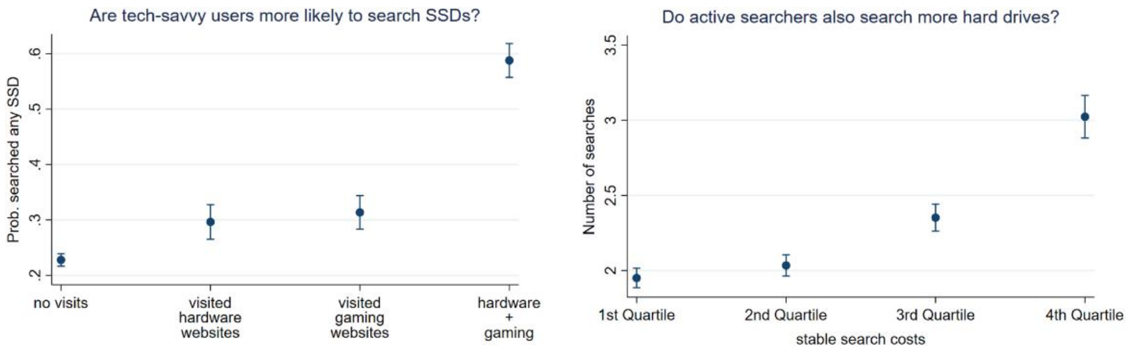
Note. The table shows the average prices of hard drives, the average standard deviation of prices across days, the number of days they were sold with at a discount (i.e., a price below median), and the average discount depth.

2.4.2. Consideration Shifters. I also construct several variables measuring users' expertise in computer hardware. To identify computer enthusiasts, I locate users who visit specialized websites about PC hardware and gamers who might naturally have more computer expertise. To this end, I identify users who visited at least one website related to PC hardware, video games, or e-sports (see Online Appendix A for details). These users, whom I refer to as *tech-savvy* users, are more knowledgeable about technology, so they are more likely to be aware of the SSD category and might better understand the benefits offered by SSDs. Consistent with this idea, Figure 1 (left panel) shows that both hardware enthusiasts and gamers are more likely to search at least one SSD during their search session than other users. Table 6 in the Online Appendix provides further evidence of this effect by regressing key outcome variables on the tech-savvy indicator (i.e., indicator that a user visited PC hardware or gaming websites at least once) and on search cost shifters described later. The results show that tech-savvy users are about 70% more likely to search at least one SSD during their search session, suggesting that they are indeed more likely to consider SSDs. Although tech-savvy users search 30% more options, most additional searches are SSDs and only a few are HDDs. Therefore, tech-savvy users search more, which mostly reflects that they examine more SSDs. One might worry that tech-savvy users have substantially different preferences, thus violating the exclusion restriction. To examine this concern, I test whether tech-savvy users are more likely to purchase expensive hard drives and hard drives with high storage capacity within the set of hard drives they search (see Table 6 in Online Appendix A.3). I cannot reject the null hypothesis that tech-savvy and non–tech-savvy users make the same purchase decisions conditional on search; therefore, I do not find any evidence that these users have substantially different preferences.

2.4.3. Search Cost Shifters. To construct search cost shifters, I study how much users search in other product categories on Amazon. While search costs may vary by context, product category, or even time of day, the search costs of individual users might be somewhat stable across categories. This stability might reflect the user's opportunity cost of time or their general ability to process information quickly and efficiently. I isolate this stable component by computing the average number of searches each user made in categories other than hard drives. In 2016, users visited about 3.5 million Amazon product pages of items other than hard drives. I classify all these 3.5 million products into categories by using category classifiers reported on Amazon product pages (see details in Online Appendix A). I then compute how many items each user searched in each category. Next, I regress the number of searches on user fixed effects and category fixed effects to partial out the impact of category-specific search costs. In what follows, I use the estimated user fixed effects to measure the so-called *stable search propensity* of this user. Figure 1 (right panel) illustrates that higher values of this search cost shifter are associated with more active search in the hard drive category on Amazon. Table 6 in the online appendix further shows that the preferences of these "active searchers" do not seem to systematically differ from those of other users.

I also attempted to measure how much time users spend on other online activities that compete for their limited attention (e.g., social media, emails and conference calls, news reading). I found that such time-use variables do not always correlate with search behavior in intuitive ways (see Online Appendix A.4 and Table 6 for details). Many of them also correlate with conditional purchase probabilities, suggesting they cannot be excluded from the preference equation. The only time use variable I include as a search cost shifter is the *daily total time online*, which is not correlated with

Figure 1. (Color online) Consideration and Search Cost Shifters Correlate with Users' Search Decisions



Notes. Both sets of shifters are introduced in Section 2.4. Online Appendix A.4 explains in detail how I identified visits to specialized hardware websites and gaming websites. Similarly, Online Appendix A.3 describes how I constructed the stable search propensity using individual-level search data from other product categories.

conditional purchase probabilities. Users who spend more time online likely have a lower opportunity cost of time; therefore, they might have lower search costs.

3. Empirical Choice Model

The user starts the session by opening the hard drive category page, which presents available hard drives and their basic attributes (top panel in Figure 2). The user can then narrow down the list of options to specific categories (e.g., SSDs). Because I do not observe which categories users consider, the model treats category consideration as an unobserved process. The user then "searches" selected hard drives within the considered categories by opening product pages and reading detailed product descriptions (bottom panel in Figure 2). I capture this process using a sequential search model. After finishing search, the user buys one of the searched hard drives or leaves the website without buying.

3.1. Utility Model

Consider a user who chooses from J available hard drives. The user i demands exactly one hard drive and derives the following indirect utility from buying a hard drive j:

$$u_{ij} = x_j'\beta_i - \alpha_i p_{j,t(i)} + \eta_{ij} + \varepsilon_{ij}, \tag{1}$$

where $p_{j,t(i)}$ is the price of hard drive j during the week of the visit $t(i)$, x_j is a vector of hard drive attributes that includes a constant, and α_i and β_i are the user's price sensitivity and tastes. Although Amazon does not personalize prices, I index price with $t(i)$ because different users visit the hard drive category on different weeks, thus getting exposed to different prices. I use this cross-sectional variation in prices to identify consideration separately from tastes, following the identification arguments outlined in Section 3.4. The specification in (1) assumes that hard drive prices do not change within the focal week $t(i)$, implying that within-week price changes are not used in estimation.

The vector x_j captures all time-invariant attributes that users observe in the category list (see the top panel of Figure 2). In my application, x_j includes the brand, type (internal versus portable), speed in megabytes per second, and storage capacity in terabytes. Additionally, Online Appendix E shows that model estimates are robust to the inclusion of hard drive quality and reliability in the utility function. I assume that the indirect utility in (1) is a linear function of attributes x_j. Relaxing this linearity assumption would be difficult because in a model with an arbitrarily flexible utility function, it would be hard to identify consideration parameters separately from preferences.[6] To ensure that preferences are identified, I also assume that both SSDs and HDDs are located in the same space of characteristics x_j. That is, I assume that prior to searching, users do not observe any other attributes not included in x_j that might discourage them from examining HDDs or SSDs. The terms η_{ij} and ε_{ij} in (1) are independent and identically distributed (i.i.d.) mean zero stochastic terms, normally distributed with variances σ_η^2 and σ_ε^2. In estimation, I normalize σ_ε^2 to fix the utility scale, and I estimate σ_η^2 together with other parameters.[7] These stochastic terms affect search decisions in different ways, as explained in Section 3.2. Finally, the user can choose the outside option that yields utility $u_{i0} = \overline{u}_0 + \varepsilon_{i0}$ with $\varepsilon_{i0} \sim N(0, \sigma_\varepsilon^2)$. I normalize the mean utility of the outside option to zero for identification, so that $\overline{u}_0 = 0$.

The terms α_i and β_i capture user i's tastes and form this user's preference profile, $\theta_i = (\alpha_i, \beta_i')'$. To simplify estimation, I assume that each element k of this profile, θ_i^k, is independent from others and follows a univariate normal distribution such that $\theta_i^k \sim N(\pi_k' w_i^p, \sigma_k^2)$, where w_i^p is a vector of taste shifters. In principle, one can specify a more flexible distribution of types θ_i, but estimating such a flexible structure would likely require more data than I have. Although this way of modeling preference heterogeneity is standard in the literature, I extend user characteristics w_i^p to include novel demand shifters that reflect consumer behavior in other product categories. In my context, taste shifters w_i^p include variables capturing brand choices of user i in other Amazon categories, this user's recent purchases of other electronics products on Amazon, and the use of file-sharing websites and cloud-storage solutions (see Section 2.4).

3.2. Information Model

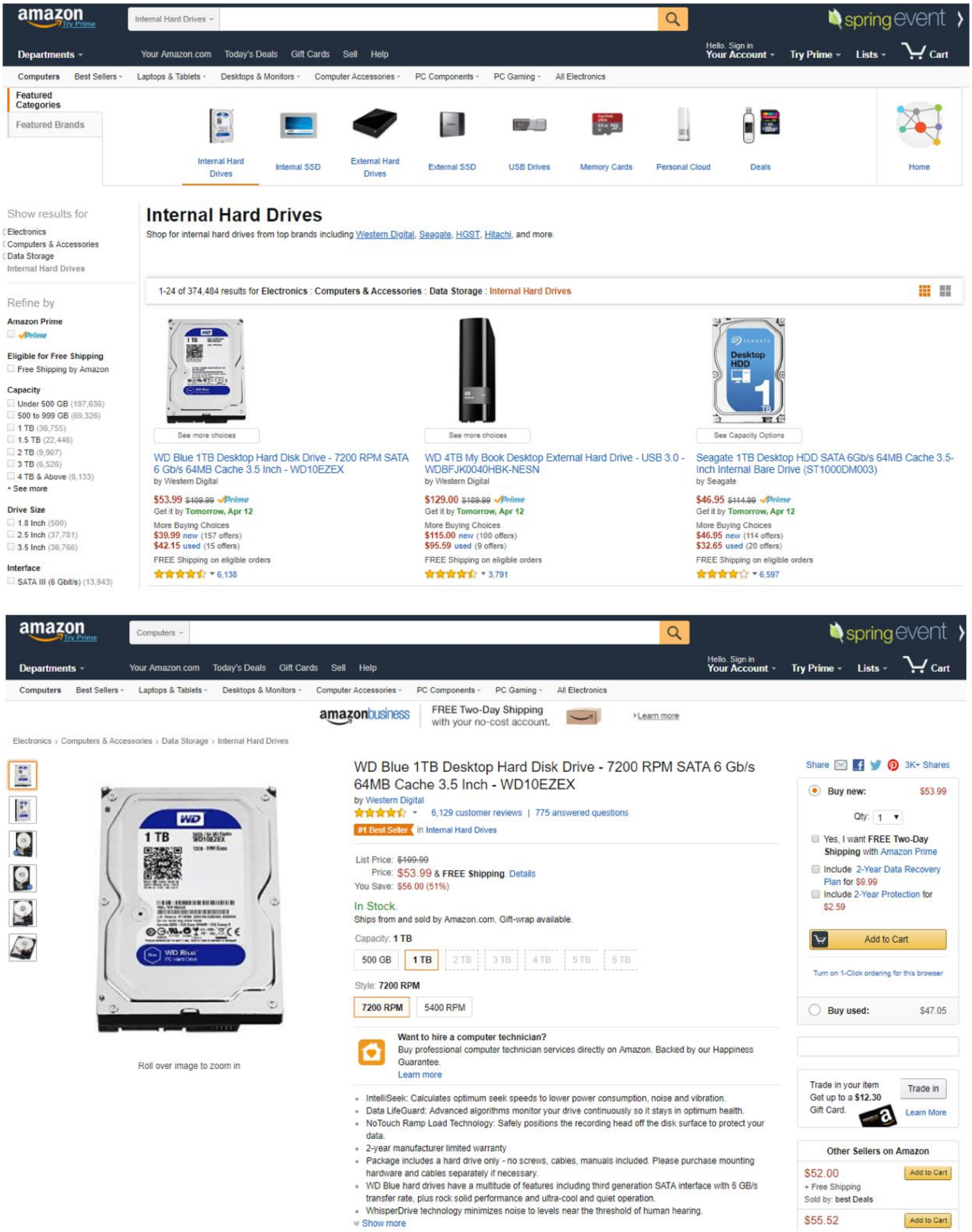
3.2.1. Category Consideration Stage. To model category consideration, I assume that the user i considers a hard drive j if and only if its *consideration index*, q_{ij}, is positive:

$$q_{ij} = d_j^{HDD}\gamma_i^{HDD} + d_j^{SSD}\gamma_i^{SSD} + \mu_{ij}, \tag{2}$$

where d_j^{HDD} and d_j^{SSD} are HDD and SSD indicators for hard drive j, γ_i^{HDD} is the propensity of user i to consider a given HDD, γ_i^{SSD} is the propensity of this user to consider a given SSD, and the term $\mu_{ij} \sim N(0, \sigma_\mu^2)$ is the i.i.d. stochastic shock. All hard drives that have positive consideration indices q_{ij} constitute this user's consideration set, $C_i \subseteq J$. I model the heterogeneity in consideration propensities γ_i^{HDD} and γ_i^{SSD} by assuming that they follow univariate normal distributions such that $\gamma_i^m \sim N(\pi_m' w_i^c, \sigma_m^2)$ with $m = HDD, SSD$, where w_i^c are consideration shifters and σ_m^2 is the variance of unobserved heterogeneity.

Although consideration propensities γ_i differ across users, these propensities are stable within each hard drive type for a given user. This specification enables me to capture several realistic behaviors. A user might consider only hard drives of one specific type, for example, only HDDs. Such a user would always consider HDDs

Figure 2. (Color online) Example of a Category Page (Top) and Product Page (Bottom) as They Appeared on Amazon.com in 2016



Show results for
‹ Electronics
‹ Computers & Accessories
‹ Data Storage
Internal Hard Drives

Refine by

Amazon Prime
□ Prime

Eligible for Free Shipping
□ Free Shipping by Amazon

Capacity
□ Under 500 GB (197,636)
□ 500 to 999 GB (69,326)
□ 1 TB (36,755)
□ 1.5 TB (22,446)
□ 2 TB (9,907)
□ 3 TB (6,526)
□ 4 TB & Above (6,133)
+ See more

Drive Size
□ 1.8 Inch (500)
□ 2.5 Inch (37,781)
□ 3.5 Inch (36,766)

Interface
□ SATA III (6 Gbit/s) (13,943)

Internal Hard Drives

Shop for internal hard drives from top brands including Western Digital, Seagate, HGST, Hitachi, and more.

1-24 of 374,484 results for **Electronics : Computers & Accessories : Data Storage : Internal Hard Drives**

See more choices
WD Blue 1TB Desktop Hard Disk Drive - 7200 RPM SATA 6 Gb/s 64MB Cache 3.5 Inch - WD10EZEX
by Western Digital
$53.99 ~~$109.99~~ Prime
Get it by **Tomorrow, Apr 12**
More Buying Choices
$39.99 new (157 offers)
$42.15 used (15 offers)
FREE Shipping on eligible orders
★★★★☆ 6,138

See more choices
WD 4TB My Book Desktop External Hard Drive - USB 3.0 - WDBFJK0040HBK-NESN
by Western Digital
$129.00 ~~$189.99~~ Prime
Get it by **Tomorrow, Apr 12**
More Buying Choices
$115.00 new (100 offers)
$95.59 used (9 offers)
FREE Shipping on eligible orders
★★★★☆ 3,791

See Capacity Options
Seagate 1TB Desktop HDD SATA 6Gb/s 64MB Cache 3.5-Inch Internal Bare Drive (ST1000DM003)
by Seagate
$46.95 ~~$114.99~~ Prime
Get it by **Tomorrow, Apr 12**
More Buying Choices
$46.95 new (114 offers)
$32.65 used (20 offers)
FREE Shipping on eligible orders
★★★★☆ 6,597

Electronics › Computers & Accessories › Data Storage › Internal Hard Drives

Roll over image to zoom in

WD Blue 1TB Desktop Hard Disk Drive - 7200 RPM SATA 6 Gb/s 64MB Cache 3.5 Inch - WD10EZEX
by Western Digital
★★★★☆ 6,129 customer reviews | 775 answered questions
#1 Best Seller in Internal Hard Drives

List Price: ~~$109.99~~
Price: $53.99 & **FREE Shipping**. Details
You Save: $56.00 (51%)

In Stock.
Ships from and sold by Amazon.com. Gift-wrap available.

Capacity: **1 TB**
500 GB | **1 TB** | 2 TB | 3 TB | 4 TB | 5 TB | 6 TB

Style: **7200 RPM**
7200 RPM | 5400 RPM

Want to hire a computer technician?
Buy professional computer technician services directly on Amazon. Backed by our Happiness Guarantee.
Learn more

- IntelliSeek: Calculates optimum seek speeds to lower power consumption, noise and vibration.
- Data LifeGuard: Advanced algorithms monitor your drive continuously so it stays in optimum health.
- NoTouch Ramp Load Technology: Safely positions the recording head off the disk surface to protect your data.
- 2-year manufacturer limited warranty
- Package includes a hard drive only - no screws, cables, manuals included. Please purchase mounting hardware and cables separately if necessary.
- WD Blue hard drives have a multitude of features including third generation SATA interface with 6 GB/s transfer rate, plus rock solid performance and ultra-cool and quiet operation.
- WhisperDrive technology minimizes noise to levels near the threshold of human hearing.

Show more

Share 3K+ Shares

● **Buy new:** $53.99
Qty: 1
□ Yes, I want **FREE Two-Day Shipping** with Amazon Prime
□ Include 2-Year Data Recovery Plan for $9.99
□ Include 2-Year Protection for $2.59
Add to Cart
Turn on 1-Click ordering for this browser

○ **Buy used:** $47.05

Trade in your item
Get up to a **$12.30** Gift Card.
Trade in
Learn More

Other Sellers on Amazon
$52.00 Add to Cart
+ Free Shipping
Sold by: **best Deals**
$55.52 Add to Cart

Source: The Wayback Machine (archive.org/web).

(e.g., $\gamma_i^{HDD} = +\infty$) and never consider SSDs (e.g., $\gamma_i^{SSD} = -\infty$), regardless of these hard drives' attributes. Similarly, a user might consider only SSDs but not HDDs ($\gamma_i^{SSD} = +\infty$ and $\gamma_i^{HDD} = -\infty$). The model also captures a continuum of cases between these two extremes. For instance, a user might be willing to consider both SSDs and HDDs but is a lot more likely to consider HDDs, which would correspond to the case of $\gamma_i^{HDD} \gg \gamma_i^{SSD}$. Apart from consideration propensities γ_i, the consideration sets are also affected by idiosyncratic shocks μ_{ij} in Equation (2). These shocks help me smoothen the consideration probabilities; therefore, they play a similar role to the presearch and postsearch utility shocks, η_{ij} and ε_{ij}. In this sense, the consideration model closely resembles the model of "soft" consideration sets in Goeree (2008).

As discussed in Section 2.4, tech-savvy users might be more informed about new technologies in this market and might therefore be more likely to consider SSDs. I capture this heterogeneity by including the tech-savvy indicator into the vector of consideration shifters w_i^c. The sign of the coefficients in π_{HDD} and π_{SSD} then determines whether tech-savvy users consider SSDs *instead* of HDDs or *in addition* to them. Importantly, Equation (2) excludes prices, thus giving me an exclusion restriction necessary to identify consideration parameters (Section 3.4 further develops this point).

One can think of several reasons why users do not consider all hard drive categories. Users may be unaware of certain categories, misunderstand the category's benefits, find the category too complex to understand (e.g., in the case of the new and relatively unknown SSD technology), or they may eliminate certain categories to simplify the choice process. The consideration model in (2) is consistent with all these causes, and it remains agnostic about which of these drive users' choices in my application. In other words, this model uses "consideration" to capture all behaviors which make consumers limit the scope of their information search to a natural set of hard drives (e.g., the subcategory of HDDs). Modeling category consideration is in itself not a new idea, and it has been extensively explored in the theoretical literature (Manzini and Mariotti 2012). Consideration models have also been shown to perform remarkably well at explaining consumer choices (Ching et al. 2009, Manzini and Mariotti 2010, Seiler 2013). Consistent with these results, I find that adding the consideration stage to the model substantially improves out-of-sample fit.

An advantage of using the tech-savvy indicator as a consideration shifter is that I can compare the behavior of users to that in the group of expert users who are more likely to be informed about the available hard drive technologies. A potential limitation, however, is that I need to assume that the tech-savvy indicator, included in the vector of consideration shifters w_i^c, does not correlate with the user's preferences (α_i, β_i). Although I provide an indirect test that supports this assumption in Section 2.4, there does not seem to be a natural way to test this assumption directly with my data. Developing such a direct test is an important venue for future research.

The proposed model views category consideration as a passive process unrelated to preferences. In principle, one could endogenize consideration by assuming that users choose which categories to consider while anticipating future search. Such a model would be substantially more complex to specify and solve, as users would need to form expectations over what hard drives they will discover and search in each category. It would also be difficult to estimate such a model with the data I have, as I do not observe the actual consideration decisions. I abstract away from these additional complexities and assume that consideration only depends on exogenous consideration shifters.[8]

3.2.2. Search Stage.

After choosing the consideration set $C_i \subseteq J$, the user proceeds to the search stage. I assume the user knows the attributes x_j, prices $p_{j,t(i)}$, and realized shocks η_{ij} for all hard drives in the consideration set C_i. That is, the user knows the first part of the utility in (1), $\delta_{ij} = x_j'\beta_i - \alpha_i p_{j,t(i)} + \eta_{ij}$, which I term *presearch utility.* One can think of a user who opens the category page of SSDs and learns the basic attributes of hard drives in that category. The shocks η_{ij} in this case capture user's preferences for any attributes that are difficult to quantify but that make certain options relatively more attractive (e.g., product photos). The user then searches hard drives one by one, revealing their ε_{ij} values that capture additional information displayed on product pages. By learning the values of ε_{ij}, the user effectively learns the realized utility $u_{ij} = \delta_{ij} + \varepsilon_{ij}$ of hard drive j, which I term the *post-search utility.*

I model search using the standard sequential search model of Weitzman (1979). Endowed with rational expectations about the distribution of ε_{ij}, the user examines hard drives one by one, revealing their ε_{ij} values and deciding at each step whether to continue searching. The first search is free, but the user incurs a fixed cost $c_i \geq 0$ for each subsequently searched hard drive. At any point during this process, the user can either purchase one of the searched options with realized utility u_{ij} or choose the outside option. The utility of the outside option, u_{i0}, is known to the user before search. Weitzman (1979) derives the optimal search behavior in this model. Define the reservation utility z_{ij} of hard drive j as a unique solution to the equation $\int_{z_{ij}} (u_{ij} - z_{ij}) dF(u_{ij} | I_{ij}) = c_i$, where $I_{ij} = \{x_j, p_{j,t(i)}, \eta_{ij}\}$ summarizes what the user knows about the drive j before searching it, and $F(u_{ij}|I_i)$ is the distribution utility given this knowledge. This reservation utility z_{ij} captures the level of utility that makes the user indifferent between terminating search and searching the hard drive j. The user searches hard drives in the order of descending reservation utilities z_{ij} within consideration set C_i. The user continues searching as long as at least one unsearched product in C_i has a reservation utility above the utility in hand. Once the user finishes searching or exhausts all search opportunities, the user chooses one of the searched options or the outside option, whichever yields higher utility.

I choose the sequential search model for several reasons. Since the model endogenizes search order, it helps me infer users' preferences from the order in which they search hard drives. Doing so helps me estimate preferences more precisely than would be possible with only data on the identities of searched products. This model also unlocks certain key counterfactual questions. For example, in Section 4.4, I ask how much surplus would SSDs generate if users could not search in a directed manner. This question would be difficult to answer in other search models which do not explicitly capture directed search. Last, this model assumes that individual tastes α_i and β_i are stable throughout the entire search session, which helps me abstract from complex models of consumer learning (Bronnenberg et al. 2016, Hodgson and Lewis 2020).

For estimation, it helps to reparametrize the model in the following way. I decompose the reservation utilities as $z_{ij} = \delta_{ij} + \xi(c_i)$, where $\xi(c_i)$ is a function that monotonically decreases in c_i and that only depends on the distribution of ε_{ij} (see the proof in Online Appendix B.1). Because the distribution of ε_{ij} is fixed by the scale normalization $\sigma_\varepsilon^2 = 1$, modeling the heterogeneity in search costs c_i is equivalent to modeling the heterogeneity in $\xi_i = \xi(c_i)$, which I term *search propensities*. Therefore, in practice, I first estimate the distribution of ξ_i and then recover the implied distribution of search costs c_i after estimation. As I show later, this alternative parametrization generates inequalities that are linear in search propensities ξ_i, which simplifies the process of taking posterior draws. Operationally, I assume that propensities ξ_i differ across users such that $\xi_i \sim N(\pi_s' w_i^s, \sigma_s^2)$, where w_i^s are search cost shifters and π_s are corresponding regression coefficients. In my application, search cost shifters w_i^s include the stable search propensity and the average daily time spent online (see Section 2.4).

3.3. Bayesian Estimation

Suppose we observe N users and know which hard drives each of them searched and in which order, as well as what they purchased. The goal of estimation is to use these individual search and purchase data to recover the unknown parameters of the model. It helps to think about this estimation problem as recovering the distribution of user types λ_i, which include the price coefficient α_i, tastes β_i, search propensities ξ_i, and consideration parameters γ_i^{SSD} and γ_i^{HHD}. Given the assumptions in Sections 3.1 and 3.2, each element λ_i^k of the user's type is independent from others and follows the normal distribution $\lambda_i^k \sim N(\pi_k' w_i^k, \sigma_k^2)$. The goal of estimation is then to recover the regression coefficients π_k, heterogeneity variances σ_k^2, and the variance of the presearch shock σ_η^2 from the data.

I estimate these parameters using a Bayesian approach. To the best of my knowledge, mine is the first paper that estimates a structural search model using Bayesian methods.[9] What makes the estimation of search models challenging is that, in most cases, the likelihood function does not admit a closed-form solution and has to be approximated. Many authors deal with this issue by using simulated likelihood methods (Honka 2014, Honka and Chintagunta 2015, Ursu 2018). However, because the likelihood of observing a specific search sequence is minuscule, one needs an extremely large number of draws to precisely approximate the likelihood. Because the resulting likelihood function is not smooth, the researcher also needs to introduce artificial smoothing (e.g., via kernel-smoothed simulator) that might bias the estimates (Chung et al. 2019). These issues would be even more troublesome in my application, where large assortments and user heterogeneity make the likelihood function even more complex and difficult to approximate.

To address these challenges, I develop a Gibbs sampler that approximates the posterior distribution of parameters using Markov chain Monte Carlo (MCMC) simulation methods. I find this approach appealing for several reasons. First, the MCMC sampler replaces the task of maximizing the likelihood with an algorithm that repeatedly draws from a series of conditional posterior distributions, thus removing the need to approximate the likelihood function. I found that in practice this makes the MCMC method more numerically stable and more robust to poorly selected starting values than likelihood-based alternatives. Second, the Bayesian approach is efficient at handling rich user heterogeneity (Rossi et al. 2012), and it provides a natural way to quantify uncertainty in finite-sample estimates without relying on asymptotic approximations.

To introduce the MCMC sampler, I first summarize the restrictions that observed searches and purchases impose on the model's parameters. Suppose the user i searches K_i hard drives and then purchases some hard drive y_i. Without loss of generality, let $1, \ldots, K_i$ be the indices reflecting the order in which the hard drives are searched, and with some abuse of notation, let $S_i = \{1, \ldots, K_i\} \cup \{0\}$ denote the resulting search set.[10] In what follows, I assume the search set S_i always includes the outside option. The observed decisions are optimal if and only if the following inequalities hold:

$$\begin{aligned}
&j \in C_i \Leftrightarrow q_{ij} \geq 0 && \forall j && \text{(consideration)}\\
&z_{i,1} \geq z_{i,2} \geq \cdots \geq z_{i,K_i} \geq z_{i,j} && \forall j \in C_i \setminus S_i && \text{(search order)}\\
&\max(u_{i,0}, u_{i,1}, \ldots, u_{i,m-1}) \leq z_{i,m} && \forall m \in \{1, \ldots, K_i\} && \\
& && && \text{(continuation)}\\
&\max(u_{i,0}, u_{i,1}, \ldots, u_{i,K_i}) \geq z_{i,j} && \forall j \in C_i \setminus S_i && \text{(stopping)}\\
&u_{i,y_i} \geq u_{i,j} && \forall j \in S_i && \text{(purchase).}
\end{aligned}$$

One can interpret these inequalities as follows. The user only considers hard drives with positive values of the consideration indices q_{ij} (consideration). Additionally, the user searches hard drives in the order of descending reservation utilities z_{ij} (search order), keeps searching until reaching the hard drive K_i (continuation), and stops after searching the hard drive K_i because the current utility exceeds reservation utilities of all unsearched hard drives in the consideration set (stopping). Finally, the user then either buys a hard drive or chooses the outside option, thus selecting an option from the set S_i (purchase).

The MCMC sampler takes sequential draws from the posterior distribution of unknown parameters while respecting this system of inequalities. To construct a quick sampler, I use the data augmentation technique and treat utilities δ_{ij} and u_{ij}, consideration indices q_{ij}, and user types λ_i as additional parameters to be estimated. The sampler then imputes the values of these additional parameters together with the structural parameters of interest. As a whole, this approach can be viewed as extending the standard hierarchical probit model by incorporating two information frictions, category consideration and costly search (Rossi et al. 2012, p. 75). In fact, my model nests the perfect information probit model with i.i.d. shocks ε_{ij} as a special case when users face zero search costs and consider all hard drives ($c_i = 0$ and $\gamma_i^{HDD} = \gamma_i^{SSD} = +\infty$ for all users i).

3.3.1. Implementation. One practical issue is how to generate draws of utilities (δ_{ij}, u_{ij}) for each user-product combination while satisfying a large number of nonlinear inequalities. I deal with this computational challenge by using two tricks. First, decomposing the reservation utilities as $z_{ij} = \delta_{ij} + \xi_i$ makes inequalities linear in all parameters, which removes the need to deal with nonlinear functions of search costs c_i. I recover the distribution of propensities ξ_i and infer the implied distribution of search costs c_i after estimation. Second, I simplify the system of inequalities, reducing it to a much simpler one. With this simplified system, drawing utilities δ_{ij} and u_{ij} and search propensities ξ_i becomes as basic as taking draws from truncated normal distributions. Because taking these draws is easy, I am able to impute unobserved utilities for all user-product combinations relatively quickly. This feature allows me to develop a practical MCMC sampler that scales well with large numbers of users and items. See Online Appendix C where I present the simplified inequalities, discuss the selection of priors, and derive all relevant posterior distributions.

I have tested the proposed MCMC sampler in several ways. I first used simulated search and purchase data to show that the MCMC sampler can successfully recover the true parameter values. To make these simulations informative, I made simulated data look similar to my actual sample in terms of the users' search and purchase behavior. Although recovering the model's parameters is of some interest on its own, my final goal is to study welfare gains from new products, which are highly nonlinear functions of estimated parameters. For this reason, I also verified that the MCMC sampler can successfully recover the change in consumer surplus from new product introductions. See Online Appendix D for details.

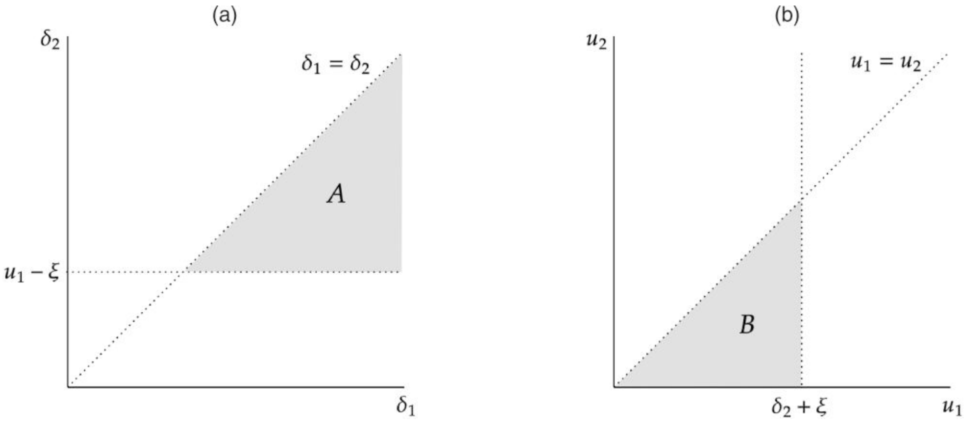
3.3.2. Illustration. Suppose a user chooses between two items indexed by $j = 1, 2$ and assume there is no outside option. Assume that in the data, a user first searches item 1, then searches item 2, and then purchases item 1. This observed search sequence imposes the following constraints on item utilities:

$$\begin{aligned}
&\delta_1 \geq \delta_2 && \text{(search order)}\\
&u_1 \leq z_2 = \delta_2 + \xi && \text{(continuation)}\\
&u_1 \geq u_2 && \text{(purchase).}
\end{aligned}$$

The first inequality holds because searching item 1 before item 2 is only optimal when the reservation utilities are ordered as $z_1 \geq z_2$, which is equivalent to $\delta_1 \geq \delta_2$ given the flat search costs. The second inequality holds because the user decided to search item 2 after learning the utility of item 1, u_1. Finally, the third inequality holds because the user purchased item 1, which must have had higher utility than item 2.

Figure 3 visualizes these inequalities. Figure 3(a) shows all permissible values of presearch utilities δ_1 and δ_2 conditional on utilities u_j and search propensity ξ. Similarly, Figure 3(b) shows all permissible values of utilities u_1, and u_2 given the values of δ_j and ξ. To implement the proposed MCMC sampling, one would need to take draws of these utilities one by one, respecting the bounds depicted in these figures and conditioning each draw on all other unknown quantities (i.e., staying within the shaded areas A and B). For example, one can first draw δ_1 from the relevant normal distribution while truncating it with a condition $\delta_1 \geq \delta_2$, then draw δ_2 truncating it with $\delta_1 \geq \delta_2 \geq u_1 - \xi$, and so on. Once utilities δ_j and u_j are drawn, one can draw the search propensity ξ, truncating it with the condition $\xi \geq u_1 - \delta_2$. The actual MCMC sampling then boils

Figure 3. Proposed MCMC Sampler



down to generating posterior draws in this fashion for each user in the data.[11]

3.4. Identification

The main challenge is how to identify consideration parameters separately from search costs and preferences. My general strategy is to rely on two distinct sets of exclusion restrictions: one that builds on the user-level shifters w_i constructed in Section 2.4 and another one that excludes prices from the category consideration decisions. First, I use the user-level shifters w_i by isolating their effect on specific stages of the choice process. For instance, I include indicators for tech-savvy users as consideration shifters w_i^c while at the same time assuming they do not impact the search process that follows consideration decisions. This approach is similar in spirit to the work by Bronnenberg et al. (2015) who compare the decisions of expert and nonexpert consumers to isolate the impact of information on choices. Similarly, I include search cost shifters w_i^s that measure the total time spent online and the average number of searches in other categories (i.e., "stable search propensity"). I assume that these variables correlate with search costs but do not directly influence the consideration decisions or the purchase decisions conditional on search. One such variable uses the extent of search in other categories as a shifter of search costs in the hard drive category, which resembles the idea of Hausman instruments that use prices in other markets as shifters of prices in a focal market (Hausman et al. 1994).

Second, I exclude prices from the consideration stage, assuming that users do not react to price changes that affect hard drives they do not consider. This assumption resembles Ching et al. (2009) who also posit that consumers do not respond to promotions when they do not consider purchasing in a given category.[12] In contrast to their panel data approach, I use cross-sectional variation in prices across users. That is, although each user makes only one purchase, I observe similar users (e.g., tech-savvy users with the same demographics) visiting Amazon on different weeks and therefore seeing different prices. To use this cross-sectional variation, I assume that the week of the visit is uncorrelated with the user's preferences, that is, tastes α_i and β_i are uncorrelated with visit time $t(i)$ in the utility model (1). I then identify category consideration from the extent to which users do not react to temporary price changes that affect specific hard drive types. For example, if exposing some of these users to lower SSD prices does not make them more likely to search SSDs, I interpret this as evidence of limited consideration.

Of course, differential responses to price changes might also be driven by heterogeneous price sensitivities α_i and tastes β_i. Because my model captures how such heterogeneity translates into choices, I use the model's structure to disentangle consideration from preferences. One may worry that, if SSD prices dropped dramatically, it would encourage many users to consider SSDs. From this perspective, it helps that virtually all variation in my data comes from temporary price discounts that reduce prices by only around 10% and are less likely to shift consideration (Table 2). An additional worry is that Amazon encourages users to consider discounted products, thus directly affecting consideration. In Online Appendix E, I show that Amazon does indeed promote discounted hard drives by placing them in more salient positions, but the effect appears to be too small to influence my qualitative results.

Another challenge is how to separate the impact of search costs from that of preferences. Because search costs c_i (and therefore search propensities ξ_i are flat across items), users must search in the order of descending presearch utilities δ_{ij} that depend only on preferences. Hard drives for which users have stronger preferences will then be searched earlier and more frequently. I can therefore identify the mean tastes $\overline{\theta}$ from the observed search order. The mean search propensity $\overline{\xi}$ can be then identified from the average number of searched hard drives.

To identify consumer heterogeneity, I mostly rely on modeling the heterogeneity in key parameters as a function of observed user characteristics w_i. These characteristics include user-level consideration and search cost shifters and other variables such preference shifters w_i^p (e.g., brand choices in other categories, searches, and purchases of electronics products) and demographics (e.g., age, income, and household size). The observed heterogeneity coefficients π_k are then identified from the extent to which users with different characteristics w_i focus their search on different subcategories of hard drives (SSDs versus HDDs), search different numbers of options, and purchase hard drive with different characteristics conditional on search sets.

I identify unobserved preference heterogeneity σ_k^2 using search data. Ideally, I would have panel data where each user searches and purchases multiple times, but such data are unavailable due to the durable nature of hard drives. Fortunately, search data provides a "mini-panel" by describing different searches made by the same user, which helps identify unobserved preference heterogeneity.[13] If all users had the same preferences, they would search hard drives, on average, in the same order. By contrast, users with heterogeneous preferences would search different sets of hard drives in the order that best matches their preferences α_i and β_i. We would then expect hard drives to have a lot more similar attributes x_j within search sets S_i of specific users than across search sets of different users. This is indeed what I see in the data, and I document this pattern for different characteristics x_j in Online Appendix A.6. These search patterns help me identify the unobserved preference heterogeneity. In turn, the unobserved heterogeneity in search propensities ξ_i is identified from the variation in the number of searches across users, beyond what is predicted by the mean search propensity $\overline{\xi}$, mean tastes $\overline{\theta}$, and unobserved preference heterogeneity parameters σ_k^2.

Finally, I need to recover the variance of the presearch shocks, σ_η^2. One can view the term η_{ij} in (1) as a structural error that determines how much hard drive prices and attributes affect search order decisions. As $\sigma_\eta^2 \to 0$, I obtain a "pure characteristic" model in which the order of search is fully driven by prices $p_{j,t(i)}$ and attributes x_j, whereas the purchase decisions conditional on search are stochastic due to the realized values of shocks ε_{ij}. By contrast, when σ_η^2 is large, both search order and conditional purchase decisions are stochastic and only weakly correlate with the hard drive attributes. Thus, I identify σ_η^2 from the extent to which the observed search order can be explained by the hard drive attributes and prices.

4. Estimation Results and Inference

4.1. Demand Estimates from the Full Model

I first estimate the complete model with category consideration and costly search. Table 3 reports parameter estimates in the form of posterior means and standard deviations, whereas the last column in Table 4 shows the dollarized values of these estimates. Most coefficients are precisely estimated and have expected signs. Users prefer faster hard drives with higher storage capacity. An average user is willing to pay a premium of \$13.7 for buying a hard drive whose average read-write speed is higher by 100 MB/s, and a premium of around \$82 for buying a hard drive with at least 1 TB of storage capacity. Users are also willing to pay substantial premia for Seagate (\$26.1), Western Digital (\$17.9), and Samsung (\$66.8), in line with the fact that these three brands attract most searches and purchases. Therefore, an average user is willing to pay about \$65 extra for SSDs because these hard drives are faster (\$48.1 premium), mostly offered by Samsung (\$66.8 premium), but often have less than 1TB of storage space (negative \$51.6 premium). The implied SSD premium of \$65 makes sense given that the average price difference between SSDs in HDDs is \$58 in the estimation sample.

Users face substantial search frictions. Both estimated information frictions, category consideration and costly search, are large in magnitude. I estimate mean consideration parameters to be $\overline{\gamma}_{HDD} = -1.106$ and $\overline{\gamma}_{SSD} = -0.917$, implying that the average user considers a given HDD with a probability 13.9% and a given SSD with a probability 18.7%. Because HDDs in this sample are a lot more numerous than SSDs, these estimates imply that most hard drives that users consider are HDDs. The left panel of Figure 4 visualizes the predicted consideration sets for non–tech-savvy users. To generate this figure, I first randomly draw 1,000 consideration propensities γ_i^{HDD} and γ_i^{SSD} from their estimated distributions, and for each drawn pair of propensities, I randomly draw 10,000 consideration sets $C_i \subseteq J$ from the consideration model in (2). Figure 4 visualizes the distribution of these randomly drawn consideration sets. We find that the average user considers only about 14 to 15 hard drives out of 100 available options, on average examining 11 HDDs and only 3 SSDs. In other words, the average user ignores about 90% of all SSDs available in this market.

I estimate the mean search propensity $\overline{\xi}$ to be 1.617. Inverting this estimate, I obtain that the implied mean search cost is \$1.5 with a standard error of \$0.18 (the formula for this inversion is derived in Online Appendix B.1). This cost is substantial given that the typical user only finds it worthwhile to search 1.7 hard drives out of 14 to 15 considered alternatives. As a result, the probability that a given user searches a specific SSD in this market is less than 2%. Although I present a more detailed welfare analysis here, these estimates do suggest

Table 3. Parameter Estimates from the Model with Search and Consideration

	Posterior			Posterior	
Coefficient	Mean	Standard error	Coefficient	Mean	Standard error
Price (100s of dollars)			Seagate brand		
Mean	−1.505	(0.050)	Mean	0.393	(0.088)
Income \$40,000–75,000	0.314	(0.100)	Searched Seagate before	0.556	(0.148)
Income \$75,000–150,000	−0.329	(0.260)	Purchased Seagate before	0.964	(0.631)
Income \$150,000+	−0.173	(0.294)	Unobserved heterogeneity	0.705	(0.047)
Unobserved heterogeneity	0.406	(0.048)			
			Western Digital brand		
Constant			Mean	0.260	(0.066)
Mean	−3.890	(0.091)	Searched WD before	0.305	(0.147)
Unobserved heterogeneity	0.532	(0.048)	Purchased WD before	0.037	(1.091)
			Unobserved heterogeneity	0.787	(0.075)
Speed (100s MB/s)					
Mean	0.207	(0.026)	Samsung brand		
Unobserved heterogeneity	0.144	(0.028)	Mean	1.005	(0.174)
			Searched Samsung before	0.145	(0.049)
Storage capacity 1–2 TB			Purchased Samsung before	−0.266	(0.698)
Mean	1.227	(0.081)	Unobserved heterogeneity	1.918	(0.321)
Uses torrents	0.276	(0.120)			
Uses cloud services	−0.021	(0.099)	Search propensity		
Unobserved heterogeneity	0.484	(0.066)	Mean	1.617	(0.041)
			Stable search propensitya	0.676	(0.218)
Storage capacity 3+ TB			Time online (hr/day)	0.017	(0.008)
Mean	1.288	(0.082)	Unobserved heterogeneity	0.469	(0.071)
Uses torrents	0.171	(0.203)	Standard deviation σ_η	1.185	(0.026)
Uses cloud services	0.516	(0.120)			
Unobserved heterogeneity	1.401	(0.124)	Consideration HDDs		
			Mean	−1.106	(0.027)
Internal drive			Tech-savvy indicator[b]	−0.182	(0.071)
Mean	−0.130	(0.093)	Unobserved heterogeneity	0.148	(0.010)
Unobserved heterogeneity	0.354	(0.092)			
			Consideration SSDs		
			Mean	−0.917	(0.070)
			Tech-savvy indicator[b]	0.778	(0.115)
			Unobserved heterogeneity	0.258	(0.024)

Notes. The table shows the estimated means and unobserved heterogeneity variances for all attributes x_j included in the model, and it additionally shows selected estimates of the observed heterogeneity. I obtain these estimates from the main sample of 1,852 users who made in total 2,872 searches. The means of all parameters correspond to the estimates for users with the average values of observed characteristics, $\overline{w}_i$. Prices are measured in hundreds of dollars, and hard drives' read-write speeds are measured in hundreds of megabytes per second.

[a] Stable search propensity is a metric capturing the average number of searches of the same user in other product categories. (see Section 2.4 for details).

[b] Tech-savvy indicator is an indicator for users who read specialized PC hardware websites or gaming news online.

that frictions prevent users from fully internalizing the benefits of the introduction of SSDs.

I estimate substantial heterogeneity in tastes, search costs, and consideration parameters. Table 3 presents the estimates of unobserved heterogeneity for all included parameters and selected estimates of the observed heterogeneity. Most importantly, these heterogeneity estimates are consistent with the empirical strategy and exclusion restrictions outlined in Sections 2.4 and 3.4, implying that additional Comscore data indeed helps me to precisely estimate preference heterogeneity and separately identify different frictions. For example, I find that brand preferences generally translate across categories. For instance, a user who searched other Seagate products in that year is also twice more likely to search a Seagate hard drive. For the other two major brands, Western Digital and Samsung, I estimate this effect to be of the same magnitude.

I also find rich heterogeneity in consideration propensities and search costs. Although the average search cost is \$1.5, it is only \$1.1 for users who spend three to four more hours online daily, and only \$0.6 for *active searchers*, that is, users who search on average two to three more products in other categories on Amazon. It is reassuring to find that the search cost shifters proposed in Section 2.4 indeed strongly correlate with the estimated search costs. I also estimate search propensities to decrease with age and increase with income, suggesting that older and poorer consumers are most affected by search frictions. Turning to consideration, I find that users identified as tech-savvy have different

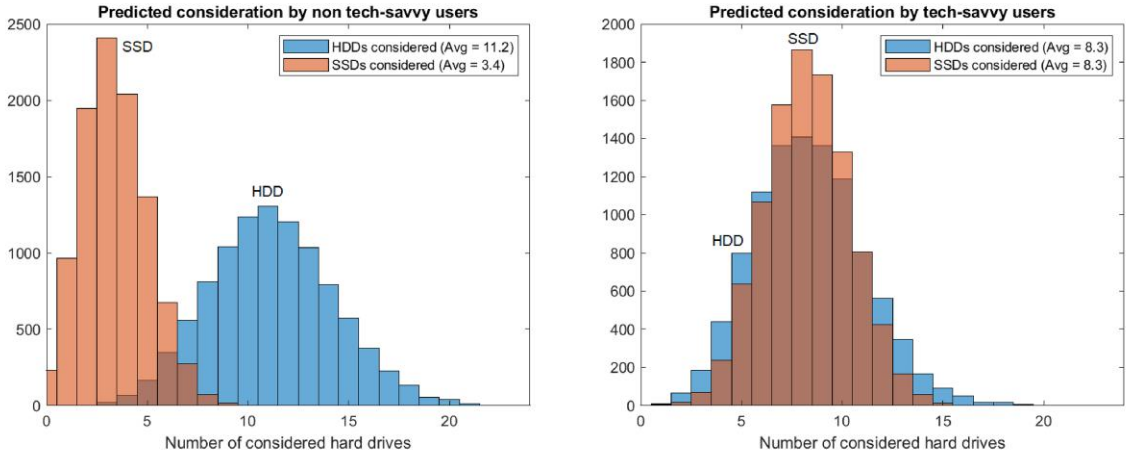
Table 4. Comparison of Preference Estimates in Models with and Without Information Frictions

	Perfect information model			Model with search and consideration		
	Posterior mean	Posterior standard error	Dollarized value	Posterior mean	Posterior standard error	Dollarized value
Mean preferences						
Price (100s of dollars)	−0.401	(0.063)	—	−1.505	(0.050)	—
Speed (100s of MB/s)	0.025	(0.009)	$6.2	0.207	(0.026)	$13.7
Internal drive	−0.168	(0.034)	−$41.9	−0.130	(0.093)	−$8.6
Storage 1 to 2 TB	−0.129	(0.017)	−$32.2	1.227	(0.081)	$81.5
Storage 3+ TB	0.119	(0.049)	$29.6	1.288	(0.082)	$85.6
Seagate brand	0.128	(0.030)	$31.9	0.393	(0.088)	$26.1
WD brand	−0.026	(0.026)	−$6.5	0.260	(0.066)	$17.3
Samsung brand	−0.140	(0.063)	−$35.1	1.005	(0.174)	$66.8
Unobserved heterogeneity						
Price (100s of dollars)	0.118	(0.028)	—	0.406	(0.048)	—
Speed (100s of MB/s)	0.026	(0.008)	$6.4	0.207	(0.026)	$9.6
Internal drive	0.121	(0.036)	$30.1	0.354	(0.092)	$23.5
Storage 1 to 2 TB	0.082	(0.032)	$20.4	0.484	(0.066)	$32.2
Storage 3+ TB	0.067	(0.022)	$16.8	1.401	(0.124)	$93.0
Seagate brand	0.102	(0.055)	$25.6	0.705	(0.047)	$46.8
WD brand	0.047	(0.026)	$11.8	0.787	(0.075)	$52.3
Samsung brand	0.112	(0.091)	$28.0	1.918	(0.321)	$127.5

Note. The table compares estimated tastes from the perfect information model which assumes zero search cost and full category consideration (columns 1-3) with those from the full model with category consideration and costly search. I obtain the estimates from the full model using the main sample of 1,852 users who made in total 2,872 searches. The estimated mean values of preferences α_i and β_i are reported for users with the average values of observed characteristics, $\overline{w}_i$. Prices are measured in hundreds of dollars.

estimated consideration parameters. I estimate their probability of considering a given SSD to be 44.7%, which is 2.5 higher than that of non–tech-savvy users (18.7%). They also seem to consider the SSD subcategory *instead*, not *in addition* to HDDs, as their probability of considering a specific HDD (10.2%) is lower than that of other users (13.9%). The right panel in Figure 4 visualizes the predicted consideration sets for tech-savvy users and shows that these users' consideration sets tend to be equally split between SSDs and HDDs. Overall, I find that tech-savvy users are more likely to consider both subcategories during their search compared with other users, which makes them more likely to purchase an SSD.

Figure 4. (Color online) Predicted Consideration Set Sizes for Tech-Savvy and Non–Tech-Savvy Users



Notes. I first randomly draw 1,000 consideration propensities γ_i^{HDD} and γ_i^{SSD} from their estimated distributions, and for each drawn pair of propensities I then randomly draw 10,000 consideration sets using model (2). The figure on the left displays the predicted distribution of consideration set sizes for non–tech-savvy users, whereas the figure on the right shows the same distribution for tech-savvy users. Tech-savvy users are those users who, according to Comscore data, read specialized websites about PC hardware and gaming. The legend reports the average number of considered HDDs and SSDs for each user group.

4.2. Estimates Under Perfect Information

The main question that motivated this paper is how researchers should estimate the value of new goods in markets with information frictions. Because I argue that it is critical to account for frictions, I need to compare my estimates with simpler models that the literature used for measuring the value of new goods. To this end, I estimate a perfect information model that is equivalent to the model in Section 3 but in which all users face zero search costs and consider both categories ($c_i = 0$ and $\gamma_i^{HDD} = \gamma_i^{SSD} = +\infty$ for all users i). This assumption essentially reduces the model to a perfect information probit model with random coefficients.[14] Because this model does not explain why consumers search, I only use purchase data for estimation.

Table 4 shows the results from this perfect information model. Compared with the search model, the perfect information model substantially underestimates users' preference for SSDs. This discrepancy is comprised of several biases. In the perfect information models, users are willing to pay only \$6.2 for buying a hard drive whose speed is 100 MB/s higher, which is much lower than the \$13.7 premium in the full model with frictions. Instead of predicting a large and positive premium for Samsung, the most popular SSD brand, the perfect information model estimates a negative premium of \$35. When put together, the perfect information estimates imply that the average user prefers HDDs over SSDs (with a negative premium of \$2), which is at odds with the positive SSD premium of \$65 that I estimated in the main specification.

The two models yield different estimates for the following reasons. Because the perfect information model assumes users perfectly observe utilities of all products, it concludes that users do not buy SSDs because they find their attributes unattractive (e.g., high speed, Samsung brand). This incorrect inference leads to an underestimation of users' preferences for SSDs. In reality, however, the low market share of SSDs can be partly explained by information frictions. Some users do not consider SSDs altogether, whereas others consider but do search them due to high search costs. The full model in Section 4.1 correctly recognizes this and returns more plausible estimates of the users' preferences for SSD.

One may also ask whether modeling category consideration is even necessary for explaining search and purchase data. My estimates suggest that modeling consideration is indeed necessary. The full model shows good out-of-sample fit, which I discuss in Online Appendix F and illustrate in Table 14. In the same online appendix, I also explore simpler variations of the same model. Notably, I explore a "search only" version of my model that assumes users consider all products but maintains the assumption of costly search. Such a model shows significantly worse out-of-sample fit, and it underpredicts consumer surplus from SSDs by around 30%. This bias arises because the model without category consideration incorrectly assumes all users consider SSDs; therefore, it erroneously interprets limited SSD searches as a signal that users value high storage space but do not value speed. For these reasons, I treat the full model as a preferred specification and use it for welfare analysis.

4.3. Welfare Estimates

To measure how much users benefit from SSDs, I would ideally observe the hard drive market before and after the SSD introduction. However, my data only cover 2016 when SSDs were already available. Therefore, I need to predict users' choices and consumer surplus in a counterfactual scenario where SSDs are not introduced. To this end, I remove SSDs from Amazon's assortment, simulate users' search and purchase decisions from the estimated model, and calculate the change in consumer surplus.

I define consumer surplus as the ex-ante expected utility u_{ij} from the purchased hard drive, where the expectation is taken with respect consideration sets C_i, search sets S_i, purchase decisions y_i, and user types λ_i. In practice, computing consumer surplus takes a lot of time because I need to approximate a high-dimensional integral using simulations. To circumvent this issue, I use the result in Choi et al. (2018) to represent my model as a discrete choice model in which users choose an option with the highest *effective utility* defined as $v_{ij} = \min(u_{ij}, z_{ij})$. This representation removes the need to repeatedly solve the optimal search problem, thus simplifying welfare computations. I then compute the expected consumer surplus from the choice set J as follows:

$$CS(J) = E\left(\max_{j\in C_i}\{v_{ij}(\theta_i, c_i)\}\right) = \int \max_{j\in C_i}\{v_{ij}(\theta_i, c_i)\}dF(C_i|\lambda_i)dF(\lambda_i|\rho), \qquad (3)$$

where J denotes the set of available hard drives (e.g., $J = HHD$ in the counterfactual scenario without SSDs); $\max_{j\in C_i}\{v_{ij}(\theta_i, c_i)\}$ captures the highest effective utility among hard drives in the consideration set C_i; $F(C_i|\lambda_i)$ denotes the distribution over potential consideration sets $C_i \subseteq J$ given the type λ_i; and ρ is a vector of all estimated parameters. I approximate this expression using simulations (see Online Appendix G for details).

Table 5 shows the welfare estimates. Columns 1 and 2 report the unconditional surplus change, whereas columns 3 and 4 report the change in consumer surplus conditional on making a purchase. The estimates from the full model in row 1 imply that consumers derive substantial surplus from the introduction of SSDs, with

Table 5. Consumer Surplus Change from the Introduction of SSDs

		Unconditional		Conditional on purchase		Sources of ΔCS	
Model	Taste Estimates	ΔCS	Percentage of price	ΔCS	Percentage of price	Tastes	Shocks η_{ij},ε_{ij}
Full model	Full Model	$3.20 ($0.60)	3.3	$25.10 ($4.60)	26.0	57.7%	42.3%
Full model	Perfect Info	$0.70 ($0.10)	0.7	$5.50 ($0.80)	5.6	58.2%	41.8%
Perfect info	Perfect Info	$1.20 ($0.20)	1.3	$9.60 ($1.80)	10.0	40.9%	59.1%

Notes. The table reports the estimated change in the expected consumer surplus (ΔCS) after adding SSDs to the choice set of users, unconditional (columns 1 and 2) and conditional on a purchase (columns 3 and 4). The value of ΔCS is computed from the Online Appendix Equation 10 as explained in Section 4.3. Consumer surplus after the introduction of SSDs corresponds to the consumer surplus predicted by the estimated model with frictions. By contrast, I compute the surplus before the introduction of SSDs by removing SSDs from the choice set and simulating users' decisions in this new environment (see Section 4.3 for details). I then compute the surplus change by taking a difference between two surplus estimates and dividing it by the average price coefficient. The last three columns further decompose this surplus change into the effects of different utility components (columns 4–6).

the average surplus change of $3.2, around 3.3% of the average hard drive price in the sample. The perfect information model, however, yields a substantially lower estimated surplus change of only $1.2 (1.3% of the price), almost three times lower than that implied by the full model. Although both estimates seem small, the average user in the estimated model purchases only with 12% probability. Therefore, the surplus change conditional on purchasing a hard drive is as high as $25.1 or 26% of the price (see columns 3 and 4).

Two main reasons explain why ignoring frictions leads to biased welfare estimates. First, when SSDs are introduced, users in the perfect information model immediately learn the shocks ε_{ij} (and therefore utilities u_{ij}) of all new products. At least one of these utility draws is likely to be more appealing than those of the HDDs. This is why the perfect information model predicts that users are more likely to buy SSDs than they would be under information frictions, thus overestimating consumer surplus. Consistent with this explanation, switching from the perfect information to the search model, while keeping preferences fixed, reduces the estimated surplus from $1.2 to $0.7 (rows 2 and 3 in Table 5).

That standard models overestimate gains from new products is a well-known result (Bajari and Benkard 2001, Petrin 2002). One could address this by correcting the variance of taste shocks (Ackerberg and Rysman 2005) or removing these shocks altogether (Berry and Pakes 2007). My model offers an alternative solution. By explicitly capturing costly search and limited consideration, the model recognizes that users only observe the utilities u_{ij} of SSDs they actually decided to consider and search, which is typically a small subset of available SSDs. In fact, in the full model, the surplus from SSDs depends on shocks ε_{ij} a lot less than in the perfect information model (see the last column of Table 5). Thus, the full model is a more realistic model of consumer behavior in that the expected utility does not overly depend on the idiosyncratic shocks ε_{ij}.

The second reason for the observed bias is tied to the estimates of preferences. The perfect information model mistakenly attributes the low market share of SSDs to users' preferences, whereas in reality, this low share is partly explained by information frictions. As a result, the perfect information model substantially underestimates the users' preference for SSDs, thus also underestimating the surplus change. I illustrate this effect in rows 1 and 2 in Table 5, which show that switching to the "right" preference estimates from the full model increases the surplus change almost five times, from $0.7 to $3.2.

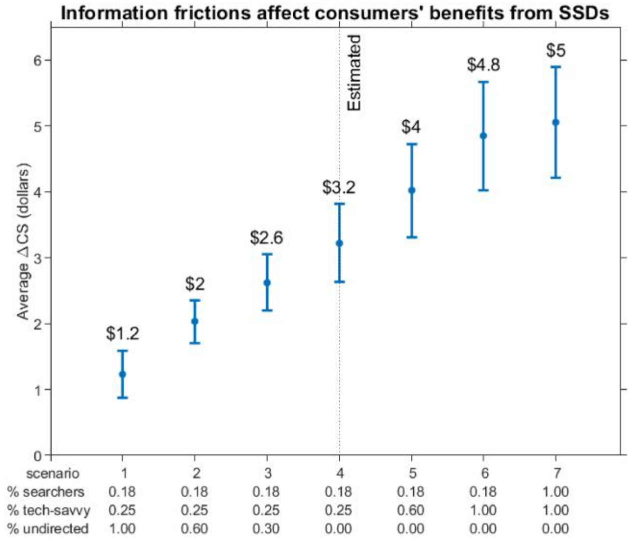
Overall, the surplus change is underestimated in the perfect information model because of implausible preference estimates. At the same time, this surplus is overestimated because the perfect information model overpredicts how many users discover an appealing SSD once this type of hard drives is introduced. In theory, the net effect of these two forces is ambiguous. Nevertheless, in my application, the model without information frictions underestimates the average consumer surplus from SSDs by a factor of three.

4.4. Information Frictions and Welfare

I have thus far argued that to estimate surplus from SSDs, one needs to account for information frictions. I now turn to a related but different question of whether frictions affect users' ability to internalize the benefits from SSDs. Answering this question is informative because firms can potentially reduce these frictions via marketing activities (e.g., advertising), and Amazon can reduce them by adopting a website design that facilitates search and encourages product exploration. The following counterfactuals explore to what extent such changes could help users to more fully internalize the surplus from SSDs.

I present two counterfactuals: one that reduces the magnitude of information frictions from their estimated level, and the other one that makes frictions more extreme. First, I gradually remove the two frictions: category consideration and costly search. Figure 5 shows the resulting change in the estimated consumer surplus from SSDs. Scenario 4 in the middle indicates the surplus

Figure 5. (Color online) Relationship Between Information Frictions and Consumer Surplus from SSDs



Notes. The graph shows the change in consumer surplus from adding SSDs to the assortment. Scenario 4 in the middle shows the surplus predicted from the estimated model. Scenarios 5 and 6 increase the number of tech-savvy users to 60% and 100%, and scenario 7 additionally reduces the search costs of all users to the level of search costs of "active searchers." Scenarios 1–3 add frictions by increasing the share of users who engage in undirected search.

change in the estimated model, which is reproduced from Table 5. In scenarios 5 and 6, I increase the number of tech-savvy users from the actual 25% in the data to counterfactual values of 60% and then to 100%. To achieve this, I randomly sample non–tech-savvy users and draw their consideration parameters from the distribution of consideration propensities γ_i^{HDD} and γ_i^{SSD} I estimated for tech-savvy users in Table 3.[15] Such change could occur if hard drive manufacturers educated users about the benefits of SSD technology or if Amazon changed the category page layout to encourage the consideration of SSDs. Because making users tech-savvy also makes them consider a lot more SSDs, the average surplus from SSDs substantially increases to $4 and $4.8 (scenarios 5 and 6).

I then additionally reduce search costs of all users to the level of "active searchers" discussed in Section 4.1, thus reducing the average search cost from $1.5 to $0.6. Although doing so increases the average surplus from SSDs to $5, the change is relatively small compared with the impact of increasing consideration. This result is intuitive: in this model, users who need a high-speed hard drive can focus on searching SSDs with the most attractive attributes x_j, so higher search costs only affect how many SSDs they will search, not whether they search them at all. From this perspective, not considering the SSD category is more damaging to their welfare than a search cost increase.

Second, I also analyze scenarios in which I make it more difficult for users to do a directed search while keeping their search costs and consideration parameters fixed at the estimated levels. This scenario mimics the time before the introduction of search engines and online platforms, in which it was more difficult for users to focus their search on hard drives with desired attributes. Operationally, I make some proportion of users search in an "undirected" way, assuming they do not observe attributes x_j and $p_{j,t(i)}$ before search but know the distribution $F(x_j, p_{i,t(i)})$ from which these attributes are drawn. In such a model users search in random order with respect to attributes x_j and $p_{i,t(i)}$, meaning they need to rely on luck or conduct extensive search in order to discover hard drives with desired attributes. Switching users from a "directed" to an "undirected" search may have complex effects on user welfare; for example, adding SSDs may actually decrease the surplus of users who only care only about HDDs, because they may now have to waste their searches on SSDs that do not match their preferences. Although the optimal search rule is still the same as under directed search, the reservation utilities, call them $\tilde{z}_{ij}$, are now determined by a different equation:

$$\int_{u_{ij} \geq \tilde{z}_{ij}} (u_{ij} - \tilde{z}_{ij}) dF(x_j, p_{i,t(i)}) dF(\varepsilon_{ij}) = c_i. \tag{4}$$

The expectation over utilities u_{ij} is now with respect to the distribution of shocks ε_{ij}, attributes x_j, and prices $p_{i,t(i)}$. Assuming rational expectations, I estimate the distribution $F(x_j, p_{i,t(i)})$ using the empirical distribution of hard drive attributes and prices in the data. As before, I compute the consumer surplus using the simplified formula in (10), except that I now use the "undirected" reservation utilities $\tilde{z}_{ij}$ and compute the effective utilities as $\tilde{v}_{ij} = \min(u_{ij}, \tilde{z}_{ij})$.

Figure 5 shows the resulting surplus estimates (scenarios 1–3). I gradually increase the number of users doing undirected search by increasing their share from zero in the benchmark scenario first to 30% in scenario 3, to 60% in scenario 2, and finally to 100% in scenario 1. The results reveal that removing directed search substantially decreases the surplus from SSDs. The surplus drops from $3.2 in the benchmark case, where all users search in a directed way, to $2.6, then to $2, and eventually decreases to $1.2 in scenario 1 where all users search in an undirected way. Overall, these results suggest that the ability to do directed search, facilitated by the existing search tools, substantially increases the ability of users to benefit from SSDs.

When put together, the scenarios in Figure 5 range from a frictionless market to a market with severe frictions where users essentially search in random order. As we move toward cases with more frictions, SSDs generate less surplus, because it becomes increasingly difficult for users to discover SSDs that match their tastes. The generated surplus can be anywhere between $1.2 and $5 depending on the magnitude of frictions. Among other things, this observation suggests that any technology shifts that reduce frictions, such as Internet penetration or new comparison platforms, may help users reap additional benefits from newly introduced hard drives. Whether such changes can, in turn, encourage manufacturers to develop and produce new hard drives is a fascinating question for future research.

5. Discussion and Conclusions

Many important economic questions hinge on the extent to which consumers benefit from new goods. By estimating benefits from new goods, researchers can provide insights about consumer surplus from product innovations (Hausman 1996, Petrin 2002), benefits from increasing product variety (Brynjolfsson et al. 2003), and gains from international trade (Broda and Weinstein 2006). This paper provides an empirical framework for estimating consumer surplus from new goods under information frictions. Using the application of hard drives, I show that accounting for frictions is critical for obtaining plausible welfare estimates. I also argue that information frictions substantially diminish consumer gains from new goods. Broadly speaking, this result implies that frictions prevent consumers from fully internalizing the surplus from new product introductions.

My analysis has several limitations. Because of data restrictions, I only observe consumer behavior on Amazon but not in other online or offline stores. All results therefore apply to Amazon consumers and cannot be immediately extrapolated to the whole population of hard drive buyers. It would be helpful to construct more complete measures of the consumer search process, documenting how consumers search across stores and channels. Doing so might generate novel research questions such as whether the competition between stores creates an environment more conducive to new product discovery. Another limitation of my data is that I do not observe whether consumers limited their search to a subcategory of products, for example, by using a search query "solid state drives." Without this additional information, I can only infer category consideration indirectly from the observed searches and purchases. Next, it would also be interesting to study the extent to which consumers do not consider SSDs because this category features relatively unknown brands. Although technically I could model consideration as a function of brand-specific variables (e.g., advertising), whether one can identify such a consideration model remains an open question. Finally, another limitation of my work is that in the model, Amazon plays a passive role in the consumer search process. In reality, Amazon may strategically design the website layout to promote certain hard drives. All analyses here should therefore be interpreted as conditional on the current website's layout.

For practical reasons, I abstracted away from several nuanced features of the search process that might characterize online shopping. Although the model assumes that search costs do not vary across hard drives, in reality, consumers may find it more difficult to gather information about certain hard drive types (SSDs). One could model such an environment by estimating product-specific search costs as in Ursu (2018). I leave this extension for future research.[16] I have also abstracted from the role of consumer learning and assumed that consumers' preferences remain stable during search. By contrast, recent empirical research shows that consumers often "zoom in" on a small set of similar products (Bronnenberg et al. 2016). Although there have been some initial attempts to model search with learning (Dzyabura and Hauser 2019, Hodgson and Lewis 2020), it would be interesting to study what such learning implies for the way consumers react to new product launches. For example, if a consumer is learning their preferences during search, they may not realize that the new product matches their preferences well; they may instead focus their initial search on products that later turn out to be irrelevant. Whether such learning behavior reduces consumer surplus from new products is an intriguing question for future work.

Acknowledgments

The author thanks advisors Bart Bronnenberg, Peter Reiss, and Stephan Seiler for exceptional supervision and constant support; Paulo Albuquerque, Kristina Brecko, Robert Donnelly, Liran Einav, Pedro Gardete, Brett Gordon, Wesley Hartmann, Avery Haviv, Han Hong, Megan Hunter, Przemyslaw Jeziorski, Anja Lambrecht, Suraj Malladi, Navid Mojir, Sridhar Moorthy, Sridhar Narayanan, Andres Perlroth, Peter Rossi, Navdeep Sahni, Paulo Somaini, Raluca Ursu, Xavier Warnes, and Ali Yurukoglu, as well as seminar participants at Stanford University, Institut Européen d'Administration des Affaires (INSEAD), Tilburg University, London Business School, Harvard University, Northwestern University, University of Washington, Rochester University, New York University, University of California Berkeley, and University of California Los Angeles (UCLA), for insightful comments and suggestions that have greatly improved the paper; Julie Williamsen, Sean Kaneshiro, and Wonhee Lee for helping me access Comscore data; and Robyn Lockwood for teaching me to write well, although it will take a lifetime to master what she taught me. All errors are my own.

Endnotes

[1] For example, Toyota spent more than $1.5 billion on advertising newly released *Corolla*, *Camry*, and *Prius* models in 2019. Disney spent more than $200 million in 2019 to advertise its new *Avengers: Endgame* movie. Apple spent $132.2 million on TV and online ads in September and October 2019 to promote the new *iPhone 11* and *Apple TV* Plus service. Source: iSpot and Kantar AdSpender databases.

[2] Western Digital surveyed several thousand participants in the United States, Spain, Germany, France, and the United Kingdom through online surveys and on-site interviews at electronics stores (Western Digital 2016).

[3] All numbers come from the Comscore data set I describe in the next section.

[4] Comscore data are depersonalized and anonymized in a privacy-compliant manner.

[5] De Los Santos et al. (2012) and De los Santos (2018) use another version of the Comscore data set in which they only observe which online stores users visit but not which product pages they visit in each store.

[6] An extreme example is a fully nonparametric model in which preferences are defined by $N \times J$ intercepts, one for each user-product pair. Such a model could perfectly rationalize any observed search behavior without the consideration stage, thus overfitting the data and making it impossible to test for limited consideration.

[7] Estimating the presearch variance σ_η^2 enables me to express the estimated search costs in dollars. I would not be able to express search costs in dollars if both variances were normalized (Morozov et al. 2021, Yavorsky et al. 2021).

[8] Honka et al. (2017) similarly interpret consideration (which they term "awareness") as a passive occurrence unrelated to preferences, whereas they view search as an active learning process through which consumers gather information about available options.

[9] Yang et al. (2015) use Bayesian methods to estimate a boundedly rational model in which consumers engage in costly information search to learn product attributes. They model myopic consumers who make each search decision as if this were their last opportunity to gather information. By contrast, I develop Bayesian estimation for a rational search model in which consumers follow a dynamically optimal search strategy (Weitzman 1979).

[10] A more correct but cumbersome notation would be to interpret the observed search sequence as a pair $\{S_i, \pi_i\}$, where $S_i \subseteq C_i$ is the set of searched hard drives, and π_i is a one-to-one mapping from S_i to the set of natural numbers $\{1, \ldots, |S_i|\}$ capturing the order in which these hard drives are searched.

[11] This sampling procedure is substantially different from the one used for estimating a full information probit model. In a full information model, one would be left with only one inequality $u_1 \geq u_2$ (item 1 is purchased) telling us that item 1 must have generated higher utility than item 2. By contrast, search data gives us a much richer system of inequalities, visualized in Figure 3.

[12] Honka et al. (2017) pursue a different strategy and identify consideration from survey data. Since my approach relies on a revealed preference argument rather than survey data, the two empirical strategies can be viewed as complementary. One could also imagine combining the two strategies, that is, treating both survey data and limited price responses as noisy signals of consideration.

[13] In this sense, the information in search data resembles second-choice data as in Berry et al. (2004). In second-choice data, the researcher knows which product a consumer would have purchased in the absence of the first choice. Similarly, in search data, the researcher knows which products a consumer searched but did not buy. If the purchased product became unavailable, the consumer would likely switch to buying another product from the same search set.

[14] In the perfect information model, I drop the term η_{ij} from indirect utility and normalize the variance of ε_{ij} to one. Because both terms are additive, normally distributed, and known to the users, dropping one of them is without any loss of generality. As usual, all estimated coefficients are then identified relative to the normalized variance of ε_{ij}.

[15] I avoid unrealistic counterfactuals that reduce the search cost to a very low value (e.g., zero) or increase the consideration probability to values close to 100%. Such counterfactuals generate unrealistic consumer behavior (e.g., users searching 70–80 hard drives per session or buying with a probability close to one) and are not informative about the potential effects of marketing policies.

[16] To implement this extension, one would need to modify the Gibbs sampler in Online Appendix C by introducing product-specific search propensities ξ_{ij}. However, one would need to develop an empirical strategy that can identify product-specific search costs separately from consideration and preferences. An interested reader may consult Morozov et al. (2021, p. 879), who show how one can identify product-specific search costs from conditional search moments.

References

Ackerberg DA, Rysman M (2005) Unobserved product differentiation in discrete-choice models: Estimating price elasticities and welfare effects. *RAND J. Econom.* 36(4):771–788.

Bajari P, Benkard CL (2001) *Discrete Choice Models as Structural Models of Demand: Some Economic Implications of Common Approaches* (Graduate School of Business, Stanford University, Stadford, CA).

Berry S, Pakes A (2007) The pure characteristics demand model. *Internat. Econom. Rev.* 48:1193–1225.

Berry S, Levinsohn J, Pakes A (2004) Differentiated products demand systems from a combination of micro and macro data: The new car market. *J. Political Econom.* 112:68–105.

Bresnahan TF (1986) Measuring the spillovers from technical advance: Mainframe computers in financial services. *Amer. Econom. Rev.* 1:742–755.

Broda C, Weinstein DE (2006) Globalization and the gains from variety. *Quart. J. Econom.* 121:541–585.

Bronnenberg BJ, Kim JB, Mela CF (2016) Zooming in on choice: How do consumers search for cameras online? *Marketing Sci.* 35:693–712.

Bronnenberg BJ, Dubé J-P, Gentzkow M, Shapiro JM (2015) Do pharmacists buy Bayer? Informed shoppers and the brand premium. *Quart. J. Econom.* 130:1669–1726.

Brynjolfsson E, Hu Y, Smith MD (2003) Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Sci.* 49:1580–1596.

Ching A, Erdem T, Keane M (2009) The price consideration model of brand choice. *J. Appl. Econometrics* 24:393–420.

Ching AT, Erdem T, Keane MP (2014) A simple method to estimate the roles of learning, inventories and category consideration in consumer choice. *J. Choice Modelling* 13:60–72.

Choi M, Dai AY, Kim K (2018) Consumer search and price competition. *Econometrica* 86:1257–1281.

Christensen CM (1993) The rigid disk drive industry: A history of commercial and technological turbulence. *Bus. History Rev.* 67:531–588.

Chung JH, Chintagunta PK, Misra S (2019) Estimation of sequential search models. Preprint, submitted May 9, https://dx.doi.org/10.2139/ssrn.3203973.

De Los Santos B (2018) Consumer search on the internet. *Internat. J. Industrial Organ.* 58:66–105.

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

Donnelly R, Kanodia A, Morozov I (2022) The long tail effect of personalized rankings. Preprint, submitted October 10, https://dx.doi.org/10.2139/ssrn.3649342.

Dzyabura D, Hauser JR (2019) Recommending products when consumers learn their preference weights. *Marketing Sci.* 38:417–441.

Goeree MS (2008) Limited information and advertising in the US personal computer industry. *Econometrica* 76:1017–1074.

Hausman J, Leonard G, Zona JD (1994) Competitive analysis with differenciated products. *Ann. Econom. Statist.* 1:159–180.

Hausman JA (1996) Valuation of new goods under perfect and imperfect competition. *The Economics of New Goods* (University of Chicago Press, Chicago), 207–248.

Hicks JR (1940) The valuation of the social income. *Economica (New Series)* 7:105–124.

Hodgson C, Lewis G (2020) You can lead a horse to water: Spatial learning and path dependence in consumer search. Preprint, submitted August 5, https://dx.doi.org/10.2139/ssrn.3667788.

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

Honka E, Chintagunta PK (2015) Simultaneous or sequential? Search strategies in the US auto insurance industry. *Marketing Sci.* 36:21–42.

Honka E, Hortaçsu A, Vitorino MA (2017) Advertising, consumer awareness, and choice: Evidence from the US banking industry. *RAND J. Econom.* 48:611–646.

Honka E, Hortaçsu A, Wildenbeest M (2019) Empirical search and consideration sets. *Handbook of the Economics of Marketing*, vol. 1 (Elsevier, New York), 193–257.

Igami M (2017) Estimating the innovators dilemma: Structural analysis of creative destruction in the hard disk drive industry, 1981–1998. *J. Political Econom.* 125:798–847.

Manzini P, Mariotti M (2010) Revealed preferences and boundedly rational choice procedures: An experiment. Working paper, University of St. Andrews and IZA.

Manzini P, Mariotti M (2012) Categorize then choose: Boundedly rational choice and welfare. *J. Eur. Econom. Assoc.* 10:1141–1165.

Moraga-González JL, Sándor Z, Wildenbeest MR (2022) Consumer search and prices in the automobile market. *Rev. Econom. Stud.*, ePub ahead of print July 28, https://doi.org/10.1093/restud/rdac047.

Morozov I, Seiler S, Dong X, Hou L (2021) Estimation of preference heterogeneity in markets with costly search. *Marketing Sci.* 40(5): 871–899.

Petrin A (2002) Quantifying the benefits of new products: The case of the minivan. *J. Political Econom.* 110:705–729.

Rossi PE, Allenby GM, McCulloch R (2012) *Bayesian Statistics and Marketing* (John Wiley & Sons, Hoboken, NJ).

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

Ursu RM (2018) The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Sci.* 37:530–552.

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

Yang L, Toubia O, De Jong MG (2015) A bounded rationality model of information search and choice in preference measurement. *J. Marketing Res.* 52:166–183.

Yavorsky D, Honka E, Chen K (2021) Consumer search in the US auto industry: The role of dealership visits. *Quant. Marketing Econom.* 19:1–52.