



# Consumer search in the U.S. auto industry: The role of dealership visits

Dan Yavorsky<sup>1</sup> · Elisabeth Honka<sup>1</sup>  · Keith Chen<sup>1</sup>

Received: 12 December 2019 / Accepted: 9 September 2020 / Published online: 26 October 2020  
© Springer Science+Business Media, LLC, part of Springer Nature 2020

## Abstract

In many markets, consumers visit stores and physically inspect products before making purchase decisions. We view the inspection of a product at a retail location as a search for product fit. We quantify the cost and benefit from searching for product fit using a discrete choice model of demand with optimal sequential search. In these models, the benefit of searching is measured by the standard deviation of the product fit and has, heretofore, been fixed to one in estimation. We show that, with an exogenous search cost shifter, both the cost *and* benefit of searching can be separately estimated. Our empirical setting is the U.S. automotive market. We assemble a unique data set containing individual-level smartphone geolocation data that inform us about dealership visits. We also obtain information on new vehicle purchases from proprietary DMV registration data. Our exogenous cost shifter is the distance a consumer must travel to visit a dealership. Our results show that the benefit provided by dealerships to consumers is substantial. Within our empirical context, failure to estimate the standard deviation of the product fit leads to biased search cost and consumer surplus estimates and to inaccurate predictions regarding consumers' number of searches and effects of at-home test drive programs.

**Keywords** Consumer search · Automotive industry

**JEL Classification** D83 · L62 · M31

---

We are extremely grateful to Peter Rossi for his thoughtful feedback and helpful suggestions. We thank Brett Hollenbeck, Randy Bucklin, Sylvia Hristakeva, Sergei Koulayev, Stephan Seiler, Marco Testoni, Raluca Ursu, and Geoff Zheng for valuable conversations. We also thank participants of the University of Houston Doctoral Symposium, the Interactive Online IO ( $IO^2$ ) seminar, and the Stanford University QME Rossi seminar as well as of seminar participants at UCLA, UC San Diego, and UC Riverside for their comments. All errors are our own.

---

✉ Dan Yavorsky  
dan.yavorsky.phd@anderson.ucla.edu

Extended author information available on the last page of the article.

## 1 Introduction

Prior to the Internet, store visits and the physical inspection of products played a prominent role during consumers' purchase process – especially for high-ticket durable goods such as automobiles, furniture or electronics. Information on prices and product features was difficult to obtain other than via store visits. During the last two decades, a wealth of online information has become available, giving consumers an alternative method to easily gather product information including prices. Nowadays, consumers in the market for a durable good still visit brick-and-mortar stores to obtain information about product attributes which are difficult to find online. These consumer-specific preferences for certain product attributes are often called “product fit” or “match value.”<sup>1</sup> An open question is how much consumers benefit from visiting brick-and-mortar stores today.

Across many product categories, consumers in 2019 visit fewer brick-and-mortar stores and buy more goods online than they did 15 years ago. However, looking at the number of physical store visits alone is not sufficient to determine the benefit from visiting brick-and-mortar stores. The reason is that the number of visits is *jointly* determined by both the cost and the expected benefit of a store visit. Thus consumers might be visiting fewer brick-and-mortar stores because of higher cost or lower expected benefit (or a combination of both). To answer our research question, the cost and benefit of store visits must be *separately* quantified. In this paper, we show that this can be achieved in a sequential search model for product fit with access to an exogenous search cost shifter.

Our empirical application is the new car market. Consumers typically visit dealerships before making a purchase. Prior to the Internet, Ratchford and Srinivasan (1993) find that consumers made, on average, 4.6 dealership visits in 1986. Ratchford et al. (2007) find that number to be 2.2 in 2002. Morton et al. (2011) indicate that consumers, on average, visit 1-2 dealerships. By 2016, this number had decreased to 1.3.<sup>2</sup> This empirical pattern in the new car market is consistent with patterns found in many other product markets.

Our unique data on the new car market come from Texas. We observe mobile device geolocation information that captures each consumer's home location and dealership visits, i.e., individual-level offline search behavior unintrusively collected in real time. These unique data stand in contrast to previous literature on offline search that neither observes the number nor the order of searches – especially in the car market (e.g., Nurski and Verboven 2016; Murry and Zhou 2020). An exception is Moraga-Gonzalez et al. (2018). However, the authors rely on aggregate moments

<sup>1</sup>Throughout this paper, we use these two terms interchangeably. In our empirical context of new car purchases, the match value captures, e.g., how spacious and comfortable a car is, how it drives, how courteous and helpful the dealership staff is, etc. to an individual consumer.

<sup>2</sup>During a 2016 FTC Workshop on auto distribution in the U.S., Peter Welch, President of the National Automobile Dealers Association (NADA), reported that “the average number of dealerships that a consumer visits before they make a purchase, an actual purchase, has gone down from 4.1 dealerships in 2005 to today, it's 1.3” ([https://www.ftc.gov/system/files/documents/public\\_events/895193/auto\\_distribution.transcript.pdf](https://www.ftc.gov/system/files/documents/public_events/895193/auto_distribution.transcript.pdf)).

of the distribution of number of searches from retrospective survey data. Our data describe both the number and order of searches at the individual-level and are unobtrusively gathered in real time while consumers engage in the actual search behavior. We combine these data with information on new car registrations, i.e., new vehicle purchases, from the Texas Department of Motor Vehicles (Texas DMV). We supplement these data with consumer and vehicle characteristics, and information on the location of auto dealerships and the brands they carry.

Importantly, these data allow us to measure the distances between each consumer's home and each searchable dealership – distance being our exogenous search cost shifter. We estimate a sequential search model for product fit à la Weitzman (1979) and parametrize search cost as a function of distance. Through simulations, we show that exogenous variation in search cost allows us to estimate the standard deviation of the match value distribution (MVSD). This parameter is a direct measure of the potential benefit from search. In prior literature, this parameter was commonly fixed to one in estimations (see also discussions on this issue in Kim et al. (2010) and Dong et al. (2020)).<sup>3</sup>

Previous sequential search literature – which mostly uses online data – has included search cost shifters in some instances. For example, Kim et al. (2010) use the number of links to the product page, Kim et al. (2017) include a high-definition dummy and product age, and Chen and Yao (2017) incorporate the slot position of the product on the webpage when modeling search cost. The difference between these papers and our approach is that our search cost shifter is exogenous, while theirs were endogenous. In fact, it is hard to imagine an exogenous search cost shifter in the online context.

We estimate the MVSD to be 8.16. This estimate is large relative to its typical normalization to one in search models for product fit and to the scale of the utility parameters (which is typically also normalized to one for identification reasons). The magnitude of the MVSD demonstrates that dealerships continue to provide substantial value to car shopping consumers today. Further, we also find that normalizing the MVSD to one has at least three implications. Note that these implications are specific to our empirical context and model. First, it results in an overestimation of the impact of distance on search cost. For example, increasing the distance to a dealership for an urban consumer from 10 to 20 miles increases estimated search cost by 48% when the MVSD is normalized to one. The same increase in distance yields an estimated search cost increase of only 19% when the MVSD is estimated.

Second, because normalization leads to incorrect estimates of the cost and benefit of search, the normalized model yields inaccurate predictions regarding the number of searches consumers engage in and incorrect estimates of consumer surplus. Normalizing the MVSD to one leads to an overprediction of the proportion of consumers who search once and an underprediction of the proportion of consumers who make

---

<sup>3</sup>As we discuss in Section 6.2, even without an exogenous search cost shifter, search cost and MVSD are separately identified by functional form. However, as discussed by previous literature (Kim et al. 2010; Dong et al. 2020) and as shown in one of our simulation studies in Section 6.3, in practice, search cost and MVSD cannot be separately estimated when only relying on this functional form identification and without an exogenous search cost shifter.

two or more searches, i.e., the long tail of the distribution of the number of searches. Estimating the MVSD allows the model to more accurately predict the distribution of the number of searches and especially its long tail.

Furthermore, for the auto industry, we find that normalizing the MVSD to one severely overstates consumer surplus. This overestimation is primarily driven by the smaller search cost estimates in a model in which the MVSD is fixed to one. We then examine how consumer surplus varies with demographics and distance. We find that demographics and distance explain a large proportion of variation in consumer surplus, namely 40%. More specifically, being older, having more kids, and a larger distance to dealerships is associated with lower consumer surplus. Living in an urban area and belonging to a racial minority is associated with higher consumer surplus. We find no association between consumer surplus and gender, income, education, and unemployment rate.

And lastly, as a result of the biased parameter estimates, managerial and policy analyses based on these estimates will be incorrect. We demonstrate this by assessing the adoption of at-home test drives. This counterfactual is inspired by a new marketing technique implemented by Hyundai in its “Hyundai Drive” program. In this program, consumers can reap the benefits of search without incurring any distance-based search costs. We find that brands with fewer dealerships benefit more from at-home test drive programs than brands with more dealerships. This is the case because consumers have to travel farther to get to those dealerships. Thus at-home test drive programs result in a bigger search cost reduction for brands with fewer dealerships, making these brands relatively more attractive to search. In Texas, brands like Ford and Chevrolet have about three to four times as many dealerships as brands like Honda, Nissan, and Toyota. Thus the three Asian car manufacturers in our data benefit more from at-home test drive programs than the two U.S. car manufacturers. This “asymmetric” effect on market shares is not driven by a change in consumer preferences, but by a change in search costs. Comparing the predictions for at-home test drive programs from a model with MVSD fixed to one to a model in which MVSD is estimated, we find the predictions to be directionally the same but their magnitude underestimated by a factor of three to four.

This paper makes two primary contributions to the literature. First, we show that the MVSD can be separately estimated from search cost with access to an exogenous search cost shifter. Similar to search models for prices in which the standard deviation of the price distribution is commonly estimated from data, we demonstrate the importance of doing the same for search models for product fit. Further, in our empirical context, failure to estimate the MVSD leads to incorrect model parameter and consumer surplus estimates, incorrect predictions regarding the distribution of the number of searches, and incorrect conclusions from policy analyses. And second, this study is among the first to provide a detailed assessment of physical (“offline”) search using individual-level observational data. Other studies of offline consumer search behavior mostly rely on path tracking information within a store, aggregate data or retrospective survey data. Other studies of consumer search using individual-level observational data do so in an online context using cookies or other web-tracking technology to assess browsing behavior. Access to data from new

technologies such as mobile device geotracking gives unintrusive access to consumers and allows researchers to observe them in their daily behavior opening doors to new research opportunities.

The outline of this paper follows: we discuss this study in the context of the relevant literature in Section 2. In Section 3, we present the data. We provide reduced-form results in support of our modeling framework in the following section. In Section 5, we formally introduce the model before discussing identification and estimation in Section 6. We present the empirical results in Section 7 and a counterfactual in Section 8. In Section 9, we discuss limitations and directions for future research. Finally, we conclude in the last section.

## 2 Relationship to existing literature

This study contributes to the streams of literature on consumer search and on the role of dealerships in the U.S. auto industry. In the following, we review the relevant literature and delineate the positioning of our research vis-à-vis the findings from extant research.

Previous literature estimating sequential search models for product fit usually utilizes data on online browsing behavior. For example, Kim et al. (2010) and Kim et al. (2017) use aggregate view-rank and purchase data for camcorders sold on Amazon.com. Koulayev (2014), Chen and Yao (2017), De los Santos and Koulayev (2017), and Ursu (2018) study different aspects of online hotel bookings. Yao et al. (2017) offer an application to television viewing, Morozov (2019) studies new product introductions in the computer hard drive market, and Gardete and Anthill (2020) investigate online browsing for used cars. In contrast to the previously mentioned papers, we focus on offline consumer search.

Because of limited data availability, studies of offline search behavior are very rare. Two exceptions are Jain et al. (2016) and Seiler and Pinna (2017).<sup>4</sup> Jain et al. (2016) assess the effects of sales assistance and search on purchase incidence and expenditure using video recordings of a retail clothing store's product display area. Seiler and Pinna (2017) measure the returns to price search by analyzing shopping cart movements in a supermarket. These two papers differ from ours in two aspects: they focus on the duration of search, i.e., time spent searching, rather than the number of searches and they study consumer search activity within a store, while we focus on search activity across stores.

Most closely related to this paper is Moraga-Gonzalez et al. (2018) who incorporate sequential search for product fit directly into the framework of Berry et al. (1995). The authors supplement aggregate car sales data from the Netherlands with moments from a survey on Dutch consumers' search behavior. Using these data,

---

<sup>4</sup>Seiler and Yao (2017) use supermarket path tracking data combined with advertising data to investigate how advertising affects consumers along their conversion funnel. Hui et al. (2009) and Hui et al. (2013) also study offline behavior in a supermarket using path tracking data. However, these two papers investigate vice versus virtue product purchases, the effects of other shoppers on shopping paths and purchases, and the effectiveness of mobile promotions and not consumer search behavior.

Moraga-Gonzalez et al. (2018) predict the searched dealership(s) and the search order, whereas we observe these decisions directly in our data.

This study also fits into the literature on car dealership locations and how they affect consumer demand and competition. Bucklin et al. (2008) show that, prior to the Internet, new car buyers were more likely to select cars whose dealer networks had shorter distances to the closest outlet and more dealers within a given radius from the buyer. Albuquerque and Bronnenberg (2012) estimate a model of supply and demand that does not incorporate consumer search behavior, but utilizes the locations of consumers and dealers (and thus the distance between them) and find that consumers have a strong disutility for travel. Palazzolo and Feinberg (2015) generalize a full information discrete choice model to incorporate the probability that an observed consideration set is optimal while flexibly permitting consideration set substitution among available options. The authors use their model together with survey data to study the impact of vehicle redesigns and recalls on consideration and purchase. Murry and Zhou (2020) study the agglomeration-competition trade-off of dealership co-location using a sequential search model and transaction-level data for new vehicles sold in Columbus, Ohio. While search behavior is modeled, the authors do not have data on the actual searches performed by consumers. In contrast to the previous literature on car dealerships, in this paper, we observe individual consumers' dealership visits and estimate a sequential search model that uses distance to dealerships as an exogenous search cost shifter.

## 3 Data

### 3.1 Data sources

We combine data from several sources for the empirical analysis. Mobile device geolocation data inform us about consumers' home locations and dealership visits. However, these data do not provide information on the purchased vehicles. Therefore we combine the dealership visit data with data on new car registrations from the Texas DMV. By combining these two data sets, we observe both the search sequence and the purchased vehicle at the individual level. We supplement these data with consumer and vehicle characteristics, and information on the location of auto dealerships and the brands they carry.

#### 3.1.1 Search data

We obtained consumer search data from Safegraph, a company that “provides high-quality location data products” by aggregating location information from various mobile device applications.<sup>5</sup> For the three-month time period from November 1, 2016 to January 31, 2017, Safegraph provided us with two types of information for each (anonymized) individual mobile device: home locations and dealership visits. The

---

<sup>5</sup><https://github.com/YalePrivacyLab/tracker-profiles/blob/master/trackers/SafeGraph.md>

home locations were generated by Safegraph's proprietary model that is based on the (im)mobility of the device, time of day, and assumed work patterns of device owners. The home locations are stored at the geohash-8 level.<sup>6</sup> The dealership visits were generated by Safegraph merging their mobile device geolocation data with their proprietary data set of U.S. dealership geospatial locations. A unique record is a mobile device at a dealership (identified by dealership name and street address) at a specific date and time. Thus, for each mobile device, we observe the visited dealerships and the order of those visits.

The dealership visit data record all dealership visits – for any reason. As a result, they may capture behavior other than new car searches. To remove errant observations, we exclude all data for a device if the device is observed (i) at any dealership more than 25 times, (ii) at the same dealership more than 10 times, or (iii) at any dealership between 12:00am and 6:00am. These criteria are applied to exclude, for example, the device of someone who works at, delivers vehicles to, or provides janitorial services for an auto dealership. In addition, we limit the dealership visit data to device owners with home locations in the state of Texas. The remaining data consist of approximately 154,000 unique mobile devices making 277,000 dealership visits to one or more of the 1,258 dealerships identified by Safegraph.

Dealership visits of consumers shopping for a new car can spread over days or even weeks. This raises the concern that vehicles purchased at the beginning of the three-month time period may have truncated search sequences. To investigate this concern, we calculate the average number of days between the first and last search for consumers whose last search occurred in November 2016 versus during December 2016 and January 2017. We find these values to be 1.1 and 10.2 days, respectively.<sup>7</sup> These numbers suggest that many search histories for consumers who made a purchase in November 2016 are truncated. As a result, we limit our analysis to only those consumers who purchased their vehicles in the latter two months (“sample period”). This reduces the sample size to approximately 127,000 consumers making 243,000 dealership visits.

### 3.1.2 Purchase data

Our second data come from the Texas DMV. We observe all first-time titled or registered vehicles in the state of Texas during (or up to 14 days after) the sample period.<sup>8</sup> The data include the Vehicle Identification Number (VIN), the registrant, the registrant's address, the date that the title or registration paperwork was processed by the state of Texas, the gross sales price before any adjustment for a trade-in vehicle, the

---

<sup>6</sup>A geohash is a public domain geocoding system that encodes location into a short string of letters and numbers. A geohash-8 is approximately 40 meters wide by 20 meters tall, or 800 square meters.

<sup>7</sup>If we exclude consumers who only made one search, the mean and median number of days between the first and last search for consumers whose last search occurred during the first month are 9.2 and 7, while the mean and median number of days between the first and last search for consumers whose last search occurred during the latter two months are 29.3 and 25.

<sup>8</sup>In a conversation with a Texas DMV employee, her experience was that title and registration paperwork is typically processed between one and fourteen days after submission to the DMV or county tax office, and that submission typically occurs immediately upon sale of a vehicle by a dealership.



VIN for the trade-in vehicle if applicable, and the name, city, and state of the previous owner. It is important to note that information on the previous owner identifies the dealership from which a vehicle was purchased (“selling dealership”). While most of this information is available to the public through a Freedom of Information Act (FOIA) request, the personally identifying information is not. It must be obtained through a special request and its use is subject to restrictions.

To focus on new retail vehicle sales for which consumer search data may be available from Safegraph, we limit the Texas DMV data to vehicles that belong to the 35 most popular brands from model years 2015–2018 with an odometer reading of fewer than 2,000 miles, a price exceeding \$5,000, and for which the date of title occurred during the sample period (or less than 14 days after its end). These criteria are used to omit used vehicles, commercial vehicles such as freight trucks, and alternative vehicles such as tractors, motor homes, and motorcycles. In addition, we constrain the Texas DMV data to only vehicles for which the registrant has a non-PO Box home address in the state of Texas; this is a necessary criterion for combining the search and purchase data. Finally, data are excluded if more than one vehicle is titled to the same individual during the sample period or if the registrant is not an individual. Approximately 195,000 vehicles meet these criteria.

Figure 1a and b each provide a map of individuals’ home locations. Figure 1a displays home locations for searchers from the Safegraph data, while Fig. 1b displays home locations of buyers from the Texas DMV data. Although the data sources are different, the figures show a similar geographic distribution of consumers across the state and major urban areas are clearly identifiable.

### 3.1.3 Dealership data

We manually prepared a third data set consisting of all auto dealerships in the state of Texas and the brands carried at each dealership. This data collection was necessary for two reasons. First, the colloquial and legal definition of a dealership vary and we required a definition compatible with the Safegraph data. We use the term and have organized the data such that a “dealership” represents a distinct geographic area of new vehicle retailing. For example, although Randall Noe Chrysler Dodge and Randall Noe Subaru are legally distinct dealerships, their showrooms share the same building, they have the same street address, and the vehicles in inventory are adjacently located on the same lot. We therefore categorize them as one dealership. And second, to make reasonable assumptions about vehicles that were searched but not purchased, as required for the structural search model to be introduced later, we require information on the set of brands carried at each dealership.

We identify 1,314 auto dealerships that carry new vehicles in the state of Texas. This number of dealerships is very similar to the number of dealerships (1,309) listed by the Texas Automotive Dealership Association (TADA) as of February 2018.<sup>9</sup> The small difference in the number of dealerships is due to (i) dealerships that operated during the sample period but closed by the 2018 TADA count and (ii) TADA and us

<sup>9</sup>[https://www.tada.org/web/Online/About.TADA/Dealer\\_Search.aspx](https://www.tada.org/web/Online/About.TADA/Dealer_Search.aspx)



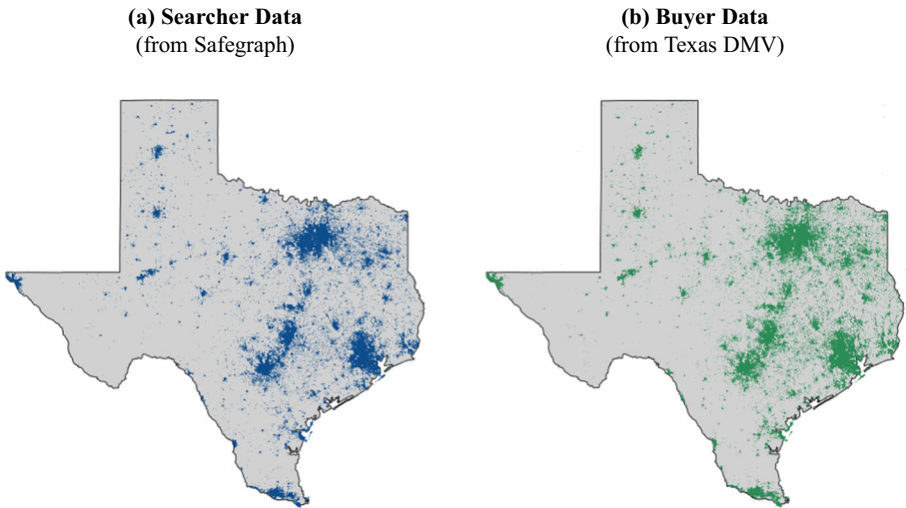


Fig. 1 Home locations

differing in how to treat two adjacent and related retail locations (TADA may recognize them as two dealerships while we count them as one or vice versa). Figure 2a plots the locations of all 1,314 auto dealerships. As expected, dealership locations are highly correlated with the geographic distribution of the population. Figure 2b provides a map of the Houston metropolitan area onto which dealership locations are overlaid. It shows that dealerships often occur spatially clustered and are generally located along major roadways.

Consistently identifying dealerships across data sets is essential for combining the search and purchase data. The manually compiled data on dealerships are matched

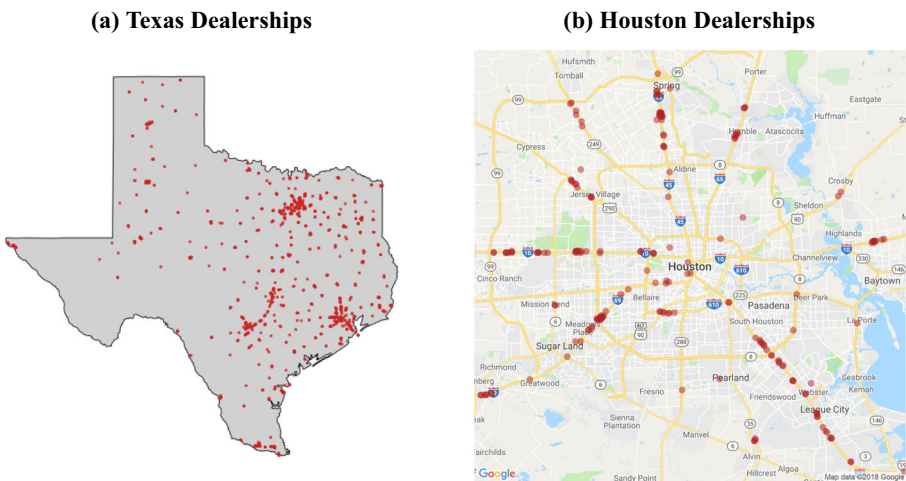


Fig. 2 Dealership locations

(i) to the dealership street addresses in the Safegraph search data and (ii) to the previous owner names in the DMV registrations data. An overlapping subset of 1,197 dealerships are identified from both data sets. Not all dealerships could be merged either because the previous owner names in the Texas DMV data were not sufficiently descriptive to uniquely identify a dealership in the Safegraph data or because the Safegraph data did not contain visits to that dealership. This latter case could occur if Safegraph did not have a geofence for that dealership or if no mobile device in the Safegraph data was observed to visit the dealership between November 2016 and January 2017.

### 3.1.4 Vehicle characteristics data

VinAudit.com, Inc., a leading vehicle data and software solutions provider for the U.S. automotive market, provided data on vehicle characteristics. The data were obtained by “decoding” the registered and trade-in VINs using the company’s API. Collected vehicle characteristics include model year, make, model, trim, “base” MSRP, vehicle type (car, SUV, truck, van), body type (e.g., sedan), number of doors, drive type (e.g., front-wheel drive), engine size, and transmission type.<sup>10</sup> We supplement these data with information on vehicle horsepower collected from Google.

We also collected data on vehicle “types” and rankings within each type from Edmunds.com. Edmunds classifies all vehicle models into one of 40 types. For example, the Honda Fit is categorized as an Extra-Small Hatchback, while the Toyota Highlander is categorized as a Midsize 3-Row SUV. Within a type, Edmunds ranks each vehicle. Their ranking process is proprietary and opaque.<sup>11</sup> The Edmunds data are used to make assumptions about (similar) searched or potentially searched, but not purchased vehicles (see Appendix A).

### 3.1.5 Consumer characteristics data

We collected consumer demographic data from the U.S. Census Bureau’s 2010 Census and American Community Survey. The data include information at the Census Blockgroup level on age, race, gender, educational attainment, income, employment status, and number of children.

<sup>10</sup>The APIs used to gather car characteristics provided “base” MSRP. The definition of “base” is manufacturer-specific and may be at either the model-level or trim-level. These base prices do not include features that customize the vehicles beyond the set of included features in the most basic configuration of the manufacturer’s model or trim. For example, if a sunroof is included as part of a specific 2017 Ford Fiesta SE but the sunroof is not a standard feature on all 2017 Ford Fiesta SEs, then the added cost of the sunroof is not included in the collected MSRP. This means that, for example, two vehicles that are largely similar (i.e., same model or same trim) but have different optional features will show the same MSRP.

<sup>11</sup>Edmunds describes its ranking process as: “Each vehicle is driven on a standardized road test loop and visits our test track for instrumented testing in controlled conditions. Our time behind the wheel is used to develop ratings that describe how a car stacks up against its direct rivals in a particular size and price class.” (<https://www.edmunds.com/new-car-ratings/>).

### 3.2 Data cleaning and construction of analysis data set

To empirically estimate a search model, we must (i) combine the search and purchase data, (ii) define the searchable set of dealerships, (iii) define the searchable set of vehicles, and (iv) create the final sample for the empirical analysis. We provide a detailed description of these four steps in Appendix A. Note that we limit the analysis sample to the five brands with the largest market shares in Texas during the sample period. This set of brands has a combined market share of 60%.

### 3.3 Descriptive statistics

The final analysis sample contains 6,511 consumers who make 7,175 visits to one of 544 dealerships. 91.1% of consumers search once, 7.8% of consumers search twice, and 1.1% of consumers search three or more times. Figure 3 shows a histogram of the number of searches. The average number of searches per consumer is 1.1. This average number of searches in our data is consistent with the most recent reports, albeit slightly smaller (see footnote 2).

Descriptive statistics on dealerships and purchased vehicles are displayed in Table 1. Chevrolet and Ford have many more dealerships than Honda, Nissan, or Toyota. The average vehicle characteristics reported in Table 1 are influenced by pickup truck sales. Chevrolet and Ford sell a higher percentage of trucks in the Texas market than the other brands and, as a result, tend to have higher average horsepower and engine sizes, while offering lower gas mileage.

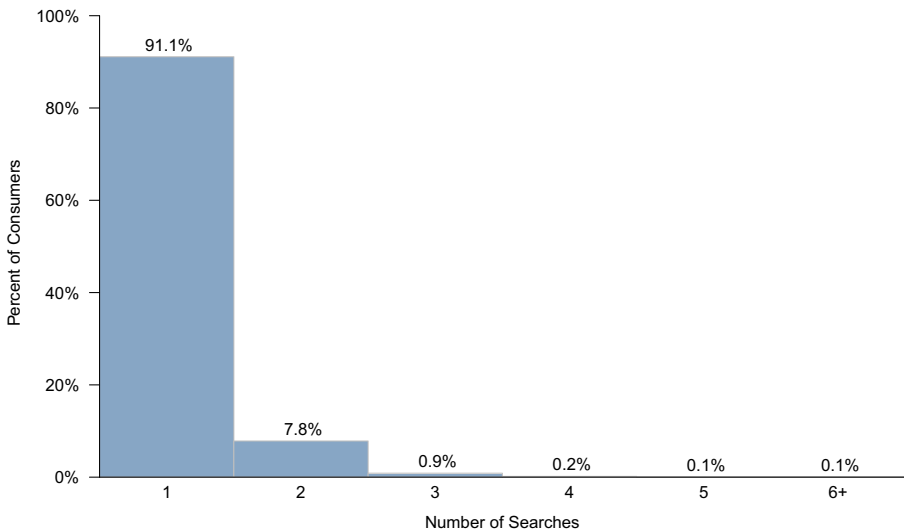


Fig. 3 Distribution of number of searches

**Table 1** Descriptive statistics

	Number of purchases in analysis sample	Number of dealerships in Texas	Mean values of purchased vehicles			
			MSRP	HP	Engine size	City MPG
Chevrolet	1,222	212	\$34,912	241	3.95	19.9
Ford	1,860	213	\$36,670	259	3.31	18.7
Honda	981	56	\$26,879	192	2.32	26.9
Nissan	652	64	\$29,944	208	2.97	23.8
Toyota	1,796	79	\$31,277	212	3.19	22.5

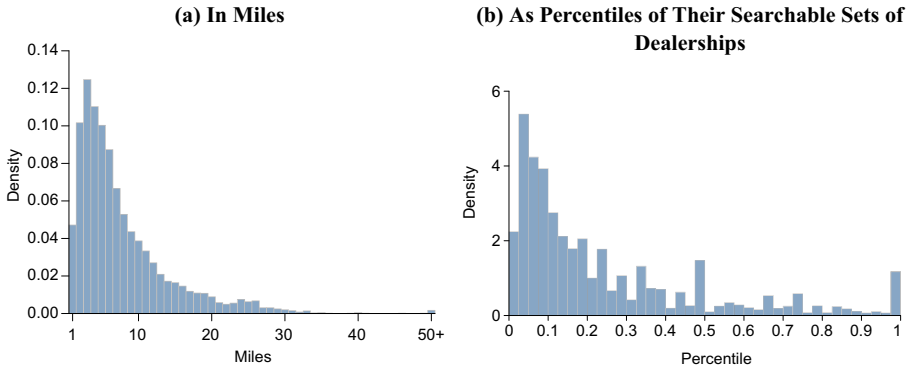
### 3.3.1 Distance to dealerships

The distance a consumer must travel to visit a dealership is our key variable. Consumers live in close proximity to many dealerships. On average, consumers in the analysis sample live within 10 miles of 6 dealerships. As we extend the radius to 20 and 30 miles, that number increases to 16 and 24 dealerships, respectively. The observation that consumers live in close proximity to many dealerships, but only search a limited number of them suggests that it is important to take their search behavior into consideration when modeling demand.

Not only are many dealerships located near consumers, but consumers also tend to purchase from nearby dealerships. The median distance from a consumer's home to the selling dealership is 5.2 miles. Figure 4a shows the distance distribution from home to the selling dealership: most consumers purchase from dealerships within 30 miles of their home.<sup>12</sup> As a point of comparison, in Fig. 4b, we display the distribution of percentiles for the distance to the selling dealership in each consumer's set of searchable dealerships. For example, suppose a consumer bought from the closest dealership and had five searchable dealerships. Then the consumer purchased from a dealership in the 0.2 percentile. The distribution in Fig. 4b shows that consumers tend to purchase from dealerships below the 0.5 percentile. The few consumers who purchase from a high percentile dealership are mostly consumers with small sets of searchable dealerships (fewer than five dealerships).

Consumers can purchase a vehicle from the closest dealership carrying the purchased brand or they can purchase from a more distant dealership carrying the same brand. 72% of consumers in the analysis sample buy from the closest dealer offering the purchased brand, the remaining 28% do not. Figure 5 shows the distance (beyond the closest dealership) to the selling dealership for those consumers who did not purchase from the closest dealership offering the ultimately purchased brand. The mean and median "extra" traveled distance are 7.6 and 5.7 miles, respectively. Buying from

<sup>12</sup>Albuquerque and Bronnenberg (2012) find the mean and median distance between consumers' homes and selling dealerships to be 10 and 7.3 miles, respectively, based on zip code centroids. These distances are also consistent with not observed, but *estimated* distance distributions reported by prior literature (e.g., Nurski and Verboven 2016; Moraga-Gonzalez et al. (2018); Murry and Zhou 2020).

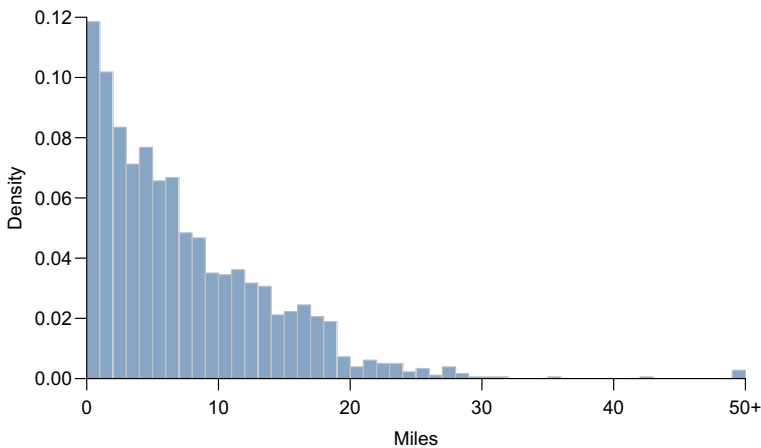


**Fig. 4** Consumers’ distances to dealerships

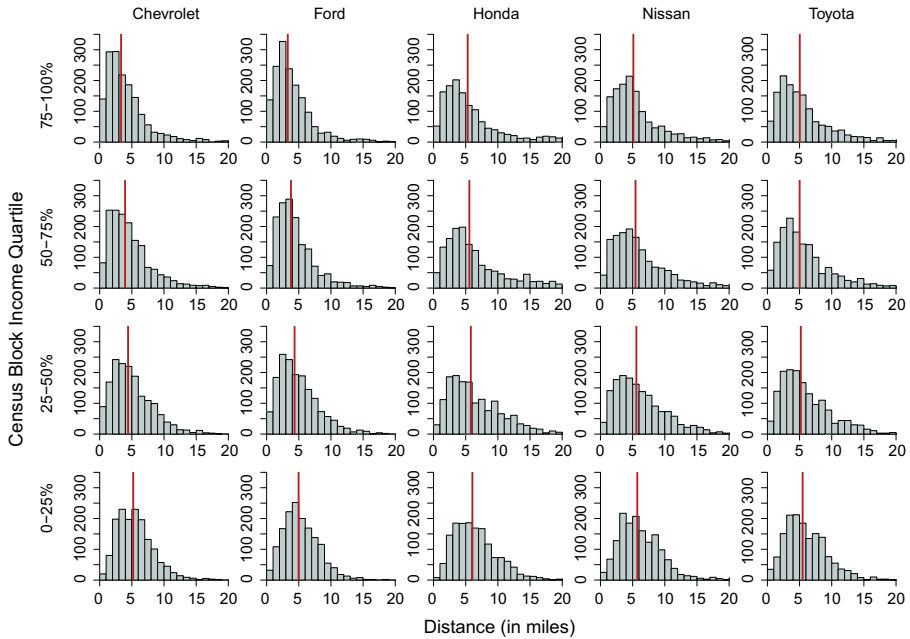
a more distant dealership suggests that that dealership offers an observed (to the consumer) expected benefit that exceeds the additional time, travel, and mental costs incurred from making the longer trip.

When modeling demand for cars, we use distance to dealerships as an exogenous search cost shifter. In the following, we provide evidence in support of our exogeneity assumption for distance. In particular, we point to three empirical patterns: first, a primary motivating factor for the choice of dealership location is accessibility. Dealerships therefore mostly locate along highways. As an example, in Fig. 2b, we show highways and dealership locations for the metropolitan area of Houston. The figure confirms that accessibility is a key driver of dealership locations. Second, previous literature (e.g., Murry and Zhou 2020) has suggested that dealerships frequently co-locate. We also find this to be the case in our data as can be seen in Fig. 2b.

And lastly, a potential concern is that dealerships might choose their locations strategically to be close to their target consumers. For example, a BMW dealer might



**Fig. 5** Miles beyond closest same-brand dealership to selling dealership



**Fig. 6** Distance histograms by brand and income

be more likely to locate in an affluent neighborhood than a Honda dealer. To investigate whether this is a concern in our data, we take the following steps: for each consumer in our final analysis sample, we calculate the distance to the closest dealership for each brand. We then split consumers into four income groups (i.e., quartiles; based on the median income in the Census Block they live in) and create histograms of the distances for each brand and income group combination. These histograms are shown in Fig. 6. The vertical lines in each histogram mark the median distance for that income group and brand combination. The histograms are of similar shape across income groups and brands and the median distances lie between 3.3 and 6.0 miles with most medians ranging from 4.3 to 6.0 miles. A potential explanation for this pattern is that the five brands we include in our empirical analysis serve a broad range of consumers.<sup>13</sup>

We conclude that, at the disaggregate level, an individual consumer's distance to dealerships can be assumed to be exogenous in our data. Other factors such as accessibility and co-location appear to drive dealership location decisions.

<sup>13</sup>We also investigated the relationship between distance to dealership and education, race, and age and found no evidence that these demographics are related to distances to dealerships (see Appendix B).

## 4 Reduced-form evidence

In this section, we investigate the set of variables that predict consumer purchase and show that including distance to dealerships is important when modeling demand for cars. We do so by estimating three standard multinomial logit models. Because consumers are assumed to choose a vehicle from a set of vehicles that are of the *same* Edmund's vehicle type, it is not necessary (or possible) to include vehicle characteristics such as the number of doors or engine type (gas vs electric) as these characteristics do not vary across vehicles within an Edmund's type. Instead, we focus on three major characteristics likely to impact consumer decision-making conditional on type: horsepower, engine size (i.e., displacement measured in liters), and estimated city mileage (MPG). In addition, because our data are from Texas, we include interactions with a dummy variable ("large vehicle") to indicate if the vehicle is a truck or a large SUV (most of which are built on a truck chassis). And lastly, we use MSRP as a measure of price.<sup>14</sup> While the data include the actual price paid for each purchased vehicle, the MSRP better reflects consumers' knowledge of prices prior to visiting a dealership. Furthermore, we observe the MSRP for the (i) purchased, (ii) searched but not purchased, and (iii) not searched vehicles.<sup>15</sup>

In Table 2, we show coefficient estimates and implied price elasticities for two full information models (a) and (b), in which the consumer is assumed to have knowledge of all searchable alternatives, and a limited information model (c), in which the consumer only chooses among the products she has searched. The two full information models differ in the included covariates with model (b) additionally including the distance to each dealership.

Across all three logit models, the coefficient estimates are similar in terms of signs and magnitudes. Coefficient significance is also generally consistent with the exception of MPG in the limited information model (c). Price negatively affects utility. On average, consumers prefer Toyota and Honda to Nissan, Ford, and Chevrolet for small and medium-sized vehicles. For large vehicles including pickup trucks, consumers prefer Ford. Among the vehicle attributes, engine size and city mileage have the expected positive sign. Horsepower is estimated to have a negative effect on utility. A potential explanation is that, conditional on vehicle type, engine size, and city mileage, vehicles configured to provide extra horsepower may require additional maintenance and are therefore not preferred by consumers.

---

<sup>14</sup>The manufacturer suggested retail price (MSRP) does not vary across dealerships (<https://www.carsdirect.com/car-pricing/what-is-the-difference-between-the-sticker-price-and-msrp>). It is set by the manufacturer and, as the name says, suggests for which price a dealer should/could sell the car. Note that sticker price and MSRP are not synonymous. The sticker price is the price a dealer lists a car for and can be lower, higher or the same as the MSRP. Sticker prices for the same (new) car can vary across dealers.

<sup>15</sup>Here, we rely on our assumption that consumers (would) search the most similar car to the one they purchase. In contrast, we do *not* observe the sticker prices (which vary across dealerships) for the (i) purchased vehicle, (ii) searched but not purchased vehicles, and (iii) not searched vehicles. While we observe the actual price paid for the purchased car, we do so after both bargaining and trade-ins. Obviously, we do not observe the actual price paid for searched, but not purchased and not searched vehicles.



**Table 2** Multinomial logit models

	Model (a)		Model (b)		Model (c)	
	Full information		Full information		Limited information	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
MSRP (in \$10,000)	-0.388**	0.034	-0.415**	0.037	-0.441**	0.169
Chevrolet	-1.786**	0.069	-1.662**	0.071	-2.658**	0.450
Ford	-1.283**	0.073	-1.104**	0.075	-2.190**	0.446
Nissan	-0.143	0.073	-0.118	0.076	-0.027	0.378
Toyota	0.069	0.055	0.188**	0.056	-0.332	0.285
Large vehicle x chevrolet	1.349**	0.115	1.375**	0.119	2.017**	0.601
Large vehicle x ford	1.824**	0.110	1.922**	0.114	2.478**	0.611
Large vehicle x toyota	-0.010	0.098	-0.064	0.101	-0.538	0.494
Horsepower (in 100)	-1.013**	0.059	-1.050**	0.061	-0.989**	0.303
Engine size	0.284**	0.034	0.324**	0.036	0.388**	0.186
MPG	0.085**	0.010	0.095**	0.011	0.027	0.053
MPG x large vehicle	0.016**	0.006	0.015**	0.006	0.034	0.031
Distance (in miles)			-0.219**	0.003		
Number of consumers	6,511		6,511		6,511	
Number of products	175,840		175,840		7,175	
Log-Likelihood	-18, 157		-13, 071		-381	
BIC	36,455		26,247		867	
<i>Own-Price Elasticities</i>						
Chevrolet	-1.18		-1.26		-0.88	
Ford	-1.21		-1.30		-0.92	
Honda	-1.00		-1.08		-0.54	
Nissan	-1.40		-1.50		-0.68	
Toyota	-1.20		-1.29		-0.81	

Note: Asterisks indicate statistical significance at the 95% confidence level. The base brand is Honda.

Recall that model (b) additionally includes distances to dealerships as a covariate. The distance coefficient is (as expected) negative, large (in absolute terms), and precisely estimated. Moreover, its addition provides substantial improvements in the log-likelihood and Bayesian Information Criterion (BIC) indicating that distance adds to the explanatory power of the model. In addition, price elasticities are more elastic when distance is included.

Comparing models (a) and (b) to model (c), price elasticities are smaller than one (in absolute terms) in the limited information model. The price elasticities are inelastic because of the large number of consideration sets of size one: consumers with such a consideration set would not select a different alternative even if prices increase because they only have one alternative available to them.

## 5 Model

### 5.1 Utility and search

We model consumer search and purchase decisions using a sequential search model for match value. A product is defined as a combination of a specific vehicle and a dealership (e.g., a Honda Civic at the John Eagle Honda Dealership of Dallas). The match value captures a mix of hard-to-quantify product characteristics that provide a unique, idiosyncratic match (or mismatch) with the consumer. The consumer learns about her match value by visiting a dealership and potentially test driving a car.<sup>16</sup> The match value might, for example, include how much a consumer likes the layout of the dashboard, how well a consumer can see in the car, how much a consumer enjoys driving the car, how courteous and helpful the dealership staff is, etc. Given that we use MSRP as the measure of price in the utility function (prior and post search), the match value also includes any deviation from MSRP due to, e.g., bargaining or a trade-in vehicle. Following previous literature models dealership visits (e.g., Moraga-Gonzaga 2018), we assume that the match value is independently and identically distributed.<sup>17</sup>

Furthermore, we make the following set of assumptions: we model consumer search conditional on choice of car type. Consumers search for the same car type (most similar car) across dealerships. And lastly, consumers learn about one car when visiting a dealership.

Consumer  $i = 1, \dots, N$  derives utility from product  $j = 1, \dots, J_i$  with utility  $u_{ij}$  given by

$$u_{ij} = \delta_{ij} + \varepsilon_{ij} \quad (1)$$

with

$$\begin{aligned} \delta_{ij} &= \mathbf{x}_j' \boldsymbol{\beta} + \eta_{ij}, \\ \eta_{ij} &\sim N(0, 1), \text{ and} \\ \varepsilon_{ij} &\sim N(0, \sigma). \end{aligned}$$

We partition what the consumer knows and does not know prior to searching. Prior to search,  $\delta_{ij}$  is known by the consumer. It is composed of a vector of observable vehicle characteristics  $\mathbf{x}_j$ , a vector of consumer preferences for those characteristics  $\boldsymbol{\beta}$ , and consumer  $i$ 's product-specific idiosyncratic preferences  $\eta_{ij}$ , which are unobserved by the researcher but known by the consumer prior to search. The second component of utility,  $\varepsilon_{ij}$ , is the product fit or match value. The consumer knows the distribution of match values,  $N(0, \sigma)$ , but she is uncertain about her specific match values  $\varepsilon_{ij}$  and must search to discover them.

<sup>16</sup>Industry reports and prior literature suggest that at least 75% of consumers test drive a car before buying it (e.g., Moraga-Gonzalez et al. (2018), <https://www.prnewswire.com/news-releases/new-research-1-in-6-car-buyers-skips-test-drive-nearly-half-visit-just-one-or-no-dealership-prior-to-purchase-255302891.html>).

<sup>17</sup>We assume that  $\varepsilon_{ij}$  is independent of  $\delta_{ij}$  as is common in the search literature (e.g., Kim et al. 2010; Chen and Yao 2017). This assumption allows us to use the three (Weitzman 1979) rules. Gardete and Anthill (2020) relax this assumption allowing for correlation between  $\varepsilon_{ij}$  and  $\delta_{ij}$ .

Search is performed sequentially and at a cost. Searching a product completely reveals the consumer's match value with that product, but does not reveal information about any other product. The cost of searching a product is parameterized as

$$c_{ij} = \exp\{\gamma_0 + \mathbf{d}'_{ij}\boldsymbol{\gamma}\}. \quad (2)$$

$\gamma_0$  is a constant and  $\mathbf{d}$  is a vector of the distance between the consumer's home location and the dealership of the searched product, an urban/rural indicator for the geographic location of the consumer, and an interaction between the latter two variables.  $\boldsymbol{\gamma}$  is a parameter vector for the search cost covariates. Consumers are assumed to have perfect recall and there is no cost for a consumer to revisit an already-searched product. Our data are conditional on search and purchase and we therefore do not model the outside option of not searching and/or not making a purchase.

## 5.2 Optimal consumer behavior

A consumer searches a product if the marginal benefit of doing so exceeds her marginal search cost. Since the match values  $\varepsilon_{ij}$  follow  $N(0, \sigma)$ , prior to search, the utilities  $u_{ij}$  follow  $N(\delta_{ij}, \sigma)$ . Define  $u_i^*$  as the highest utility among the searched products thus far.<sup>18</sup> Conditional on  $u_i^*$ , a consumer's expected marginal benefit from searching product  $j$  is given by

$$\begin{aligned} B_{ij} &= \int_{u_i^*}^{\infty} (u_{ij} - u_i^*) f_{u_{ij}}(u_{ij}) du_{ij} \\ &= \Pr[\varepsilon_{ij} > u_i^* - \delta_{ij}] \times \mathbb{E}[\varepsilon_{ij} - (u_i^* - \delta_{ij}) \mid \varepsilon_{ij} > u_i^* - \delta_{ij}]. \end{aligned} \quad (3)$$

This marginal benefit is the probability that the realized utility for product  $j$ ,  $u_{ij}$ , exceeds the best realized utility among the already searched products,  $u_i^*$ , multiplied by the expected value of  $u_{ij}$  given  $u_{ij} > u_i^*$ . As shown in Eq. 3, the marginal benefit depends on the MVSD through the integration over the utility distribution  $f_{u_{ij}}(u_{ij})$ . Holding everything else constant, a (symmetric) distribution with larger variance has more mass in the tails of the distribution and thus has both a higher probability that the next realized utility will exceed the currently best realized utility and a larger conditional expected value. Therefore quantifying the MVSD informs us about the magnitude of the marginal benefit from searching.

Weitzman (1979) derived the rules for optimal behavior under sequential search. The rules involve "reservation utilities"  $z_{ij}$ , which are the utilities that equate the marginal cost and expected marginal benefit of search. Kim et al. (2010) show that there is a closed-form solution for calculating reservation utilities  $z_{ij}$  under the assumption of normally distributed match values:

$$z_{ij} = \delta_{ij} + \zeta_{ij} \times \sigma \quad (4)$$

with  $\zeta_{ij}$  coming from the implicit function

$$\frac{c_{ij}}{\sigma} = \phi(\zeta_{ij}) - \zeta_{ij} \times (1 - \Phi(\zeta_{ij})). \quad (5)$$

<sup>18</sup>For the first search, we set  $u_i^* = -\infty$  since our data are conditional on search.

Note that  $\zeta_{ij} = \frac{z_{ij} - \delta_{ij}}{\sigma}$  such that  $\sigma$  appears on both the left- and right-hand sides of Eq. 5. We follow Kim et al. (2010)'s approach for calculating reservation utilities.

Next, we formally state Weitzman's (1979) rules. Because the rank of the reservation utilities is a one-to-one mapping with the product index  $j$ , we cast the model using  $j$  as the order of the reservation utilities such that  $j = 1$  is the product with the highest reservation utility for the consumer and  $j = J_i$  for the product with the lowest reservation utility. Let us denote the number of searches made by a consumer as  $K_i$ . For notational simplicity, we drop the consumer-specific subscript  $i$  while specifying Weitzman's (1979) rules.

Three rules govern consumer search and purchase behavior:

1. *Selection Rule:* A consumer searches products in a decreasing order of reservation utilities, i.e.,

$$z_1 \geq z_2 \geq \dots \geq z_K \geq \max_{l > K} \{z_l\}. \tag{6}$$

2. *Stopping Rule:* A consumer stops searching when the maximum realized utility among the searched products is larger than the maximum reservation utility among the unsearched products, i.e.,

$$\max_{h \leq K} \{u_h\} \geq \max_{l > K} \{z_l\}. \tag{7}$$

Equivalently, at each step during the search process, when a consumer decides to continue searching, the opposite of the stopping rule must hold, i.e.,

$$\max_{h < k} \{u_h\} < z_k \quad \forall k = 2, \dots, K. \tag{8}$$

3. *Choice Rule:* A consumer purchases the alternative with the highest realized utility among those searched, i.e.,

$$u_{j^*} = \arg \max_{h \leq K} \{u_h\}. \tag{9}$$

## 6 Estimation

### 6.1 Likelihood function

The probability that a specific sequence of searches and an ultimate purchase are made by a consumer is the probability that each of the Weitzman (1979) rules holds

at their respective steps in the consumer's search and purchase process, i.e.,

$$\begin{aligned}
 & L_i(\beta, \gamma_0, \boldsymbol{\gamma}, \sigma; \mathbf{x}, \mathbf{d}) \\
 &= \text{Prob}[\text{Consumer } i \text{ selects } j = 1 \text{ on the first search,} \\
 &\quad \text{Continues to search and selects } j = k \text{ on the } k^{\text{th}} \text{ search for } k = 2, \dots, K_i, \\
 &\quad \text{Stops searching after the } K_i^{\text{th}} \text{ search, and purchases } j_i^*] \\
 &= \int \mathbb{1}[z_{ij} \geq \max_{h < j} \{u_{ih}\} \text{ for } j = 2, \dots, K_i \cap z_{ij} = \arg \max_{k > j} \{z_{ik}\} \text{ for } j = 1, \dots, K_i \cap \\
 &\quad \max_{h \leq K_i} \{u_{ih}\} \geq \max_{k > K_i} \{z_{ik}\} \cap u_{ij_i^*} = \arg \max_{h \leq K_i} \{u_{ih}\}] dF(\eta, \varepsilon).
 \end{aligned} \tag{10}$$

The model likelihood is the product of the  $N$  individual likelihoods, i.e.,

$$L(\beta, \gamma_0, \boldsymbol{\gamma}, \sigma; \mathbf{x}, \mathbf{d}) = \prod_{i=1}^N L_i(\beta, \gamma_0, \boldsymbol{\gamma}, \sigma; \mathbf{x}, \mathbf{d}). \tag{11}$$

Note that we parametrize the MVSD as  $\theta = \log(\sigma)$  in all estimations. Neither the search nor purchase probabilities can be expressed in closed form. We approximate the integrals in the likelihood function with averages using logit-smoothed accept-reject simulation. This simulated maximum likelihood estimation algorithm follows Train (2009). Details are provided in Appendix C.

## 6.2 Identification

The parameters to be estimated include the preference parameters  $\beta$ , search cost intercept  $\gamma_0$ , parameters for the search cost covariates  $\boldsymbol{\gamma}$ , and the MVSD  $\sigma$ .

The preference parameters  $\beta$  are identified by purchase frequency, search order, and search frequency. In much the same way that the purchase decision among a set of products identifies preference parameters in a traditional discrete choice model, the purchase decision among searched products identifies preference parameters in a search model. In addition, because consumers search in order of decreasing reservation utilities  $z_{ij} = \delta_{ij} + \zeta_{ij} \times \sigma$ , holding everything else constant, products with higher  $\delta_{ij}$  ( $= \mathbf{x}_j \beta + \eta_{ij}$ ) have higher reservation utility values and are searched earlier and more frequently.

With *no* exogenous search cost shifter (or other covariates) in our model, search costs are fixed, i.e.,  $c = \gamma_0$ . In such a case, the ratio of  $\frac{c}{\sigma}$  is identified by variation in the number of searches. Holding everything else constant, consumers search little when  $\frac{c}{\sigma}$  is large and a lot when  $\frac{c}{\sigma}$  is small. Variation in the search order does not contribute to the identification of  $\frac{c}{\sigma}$ . To see this, recall that consumers search in decreasing order of reservation utilities which are a function of search cost and the MVSD  $\sigma$  (see Eq. 4). If there is a common search cost across products and a common MVSD, then  $\zeta_{ij}$  is the same for all products and the search order is entirely driven by  $\delta_{ij}$ .

Even without an exogenous search cost shifter, search cost  $c$  and MVSD  $\sigma$  are *separately* identified by parametric form. This can be seen in the implicit function  $\frac{c_{ij}}{\sigma} = \phi(\zeta_{ij}) - \zeta_{ij} \times (1 - \Phi(\zeta_{ij}))$  (Eq. 5). If  $\sigma$  were only part of the left-hand side of Eq. 5, then  $c$  and  $\sigma$  would not be separately identified, i.e., only their ratio would

be parametrically identified. However,  $\sigma$  is also part of the right-hand side of Eq. 5 – note that  $\zeta_{ij} = \frac{z_{ij} - \delta_{ij}}{\sigma}$  – ensuring parametric identification. However, in practice, search cost  $c$  and MVSD  $\sigma$  cannot both be estimated as discussed by previous literature (Kim et al. 2010; Dong et al. 2020) and as illustrated in a simulation study in the following section. Thus previous empirical research has commonly fixed  $\sigma$  to one and only estimated  $c$  (e.g., Kim et al. 2010; Chen and Yao 2017; Honka and Chintagunta 2017; Kim 2017; Ursu 2018).

However, *with* an exogenous product-specific search cost shifter, i.e., distance to dealerships in our empirical application, variation in this exogenous search cost shifter allows us to separately estimate both search cost  $c_{ij}$  (and its governing parameters  $\boldsymbol{\gamma}$ ) and the MVSD  $\sigma$ .<sup>19</sup> To be precise, all parameters in the search cost function but the search cost intercept are separately identified from the MVSD by variation in data and functional form. Even after the inclusion of an exogenous search cost shifter, the search cost intercept and MVSD continue to only be separately identified by functional form. However, we find that the inclusion of a search cost shifter helps greatly with the estimation of the search cost intercept and MVSD as we show through two simulation studies in the following section.

The reason that the parameters for the search cost covariates  $\boldsymbol{\gamma}$  and the MVSD are separately identified by variation in data is that both the number of searches and the search order inform  $c_{ij}$  and  $\sigma$  which, in turn, influence reservation utilities (see Eq. 4). For example, higher search costs yield lower  $\zeta_{ij}$  values. Thus a consumer who incurs higher cost to search a product will assign it a lower reservation utility, rank it lower in the search order, and stop her search earlier (on average) than an otherwise identical consumer facing an identical searchable set of products but with lower search cost. Thus the observable differences in search cost due to the exogenous search cost shifter coupled with the patterns of search order and search length identify the search cost parameters  $\boldsymbol{\gamma}$ .

And lastly, a higher MVSD  $\sigma$  assigns a larger value to the second component of the reservation utility (see Eq. 4). As a result, the extent to which the search order is driven by  $\zeta_{ij} \times \sigma$  rather than  $\delta_{ij}$  identifies the MVSD. In addition, with a larger MVSD (compared to a smaller MVSD) search terminates later because, holding everything else constant, a larger MVSD results in a higher marginal benefit from an additional search (see Eq. 3) and thus higher reservation utilities increasing the probability that consumers continue searching. Thus, across consumers, the average number of searches increases and the predicted distribution of the number of searches has a longer tail.

### 6.3 Simulation studies

In this section, we describe the results from three simulation studies. In the first study, we show that search cost and the MVSD cannot both be recovered without an exogenous search cost shifter. In the second study, we demonstrate that our estimation approach recovers preference and cost parameters as well as the MVSD parameter

<sup>19</sup>The exogenous search cost shifter has to be product-specific to affect the search order.

if an exogenous search cost shifter is included in the estimation. In the third study, we show that, as MVSD increases, the average number of searches increases and the distribution of the number of searches has a longer tail.

For the first and second simulation studies, we generate 5,000 consumers each searching up to five brands. The number of searchable dealerships is drawn from a combination of a chi-squared distribution and an exponential distribution so as to generally mimic the observed searchable set sizes in the empirical application. These simulated searchable set sizes range from 1 to 40 with a median of 11 and a mean of 12. Observable characteristics include the brand as well as price and city mileage; the latter two are drawn from uniform distributions on the interval  $-2$  to  $2$ . In the first simulation study, consumers have fixed search cost of  $c = \exp(0)$ . In the second simulation study, consumer search cost are additionally a function of distance to dealerships, i.e.,  $c = \exp(0 + 0.3 \cdot \text{distance})$ . Distances to dealerships are drawn from the absolute value of the sum of a chi-squared distribution with 8 degrees of freedom and a normal distribution with mean zero and standard deviation 16. Simulated distances to the searchable alternatives range from 0.01 miles to 102.4 miles with a median of 15.5 and a mean of 18.2 miles. The distributions of the number of searches in both simulation studies are generally similar to the distribution of the number of searches in our empirical application. For example, in the second simulation study, 87.7% of consumers search once, 11.4% of consumers search twice, 0.9% of consumers searches three times, and 0.02% of consumers search four times.<sup>20</sup>

For estimation, we use 1,000 draws from the distributions of the consumers' idiosyncratic preferences and match-value terms and smoothing parameters of (15, 15, 15, 5). We replicate each estimation 50 times and report mean coefficient estimates, their standard deviations, and average standard errors. The results for the first and second simulation study are shown in Table 3. The top half of Table 3 displays the results for the first simulation study in which consumers have fixed search cost, i.e., without an exogenous search cost shifter. While the preference parameters are recovered well, the average estimates of the search cost intercept and  $\log(\sigma)$  are far from their true values. This simulation study illustrates that the functional form identification is not enough to separately estimate search cost and MVSD without an exogenous search cost shifter as discussed in the previous section. The bottom half of Table 3 shows the results for the second simulation study in which consumer search cost are additionally a function of distance, i.e., with an exogenous search cost shifter. All parameters are recovered well. Importantly, the true values of the search cost intercept and MVSD, which continue to only be separately identified by functional form, can be recovered after the inclusion of an exogenous search cost shifter. Note that some of the average parameter estimates are not within two standard errors of their true values. This is not uncommon for search models estimated using SMLE (see, e.g., Honka 2014; Ursu 2018; Ursu et al. 2020).<sup>21</sup>

<sup>20</sup>In the first simulation study, 70.1% of consumers search once, 21.3% of consumers search twice, 6.3% of consumers searches three times, 1.7% of consumers search four times, and 0.5% of consumers search five or six times.

<sup>21</sup>It is well-known that estimates obtained using a kernel-smoothed frequency estimator are only asymptotically unbiased (McFadden 1989; Train 2009). The bias decreases as the scaling factor in the



**Table 3** Simulation study

	True values	Average estimate	Standard deviation of estimates	Average asymptotic standard error
No exogenous cost shifter				
Brand 1	0.2	0.177	0.055	0.027
Brand 2	0.4	0.344	0.057	0.030
Brand 3	0.6	0.522	0.055	0.025
Brand 4	0.8	0.689	0.056	0.027
MSRP	-0.5	-0.429	0.016	0.008
MPG	0.5	0.428	0.014	0.008
Cost Intercept	0	-20.900	6.499	10.983
Log Sigma	0.69	-3.976	0.309	0.386
With exogenous cost shifter				
Brand 1	0.2	0.178	0.088	0.038
Brand 2	0.4	0.374	0.064	0.045
Brand 3	0.6	0.540	0.066	0.042
Brand 4	0.8	0.714	0.080	0.038
MSRP	-0.5	-0.443	0.023	0.012
MPG	0.5	0.445	0.023	0.011
Cost intercept	-2.0	-2.623	0.167	0.078
Distance	0.3	0.347	0.015	0.007
Log sigma	0.69	0.439	0.059	0.024

For the third simulation study, we generate data sets with the same characteristics as in the second simulation study but with one exception: the MVSD. Recall that the MVSD is parametrized as  $\theta = \log(\sigma)$ . In this simulation study, we set MVSD to five different values:  $\theta = \log(2)$ , 2, 5, 10, and 25, i.e., we generate five data sets which only vary by the MVSD. For each data set, we simulate the number of searches consumers make when optimally searching sequentially using the Weitzman (1979) rules. We replicate this procedure 50 times.

The five distributions of the number of searches are shown in Fig. 7. The black lines denote the average number of searches.<sup>22</sup> As expected, they show that, as MVSD increases, the average number of searches increases and the distribution of the number of searches has a longer tail.

logit-smoothed function increases (see Web Appendix C for our definition of the logit-smoothed function), i.e., as the steepness of the logit function increases. However, increasing the scaling factor has to be balanced with the loglikelihood function retaining its smoothness and differentiability for estimation (McFadden 1989; Train 2009).

<sup>22</sup>The distributions shown in Fig. 7 are the average distributions across the 50 replications. The patterns that the average number of searches increases and that the distribution of searches has a longer tail as MVSD increases hold both on average as well as for each replication.

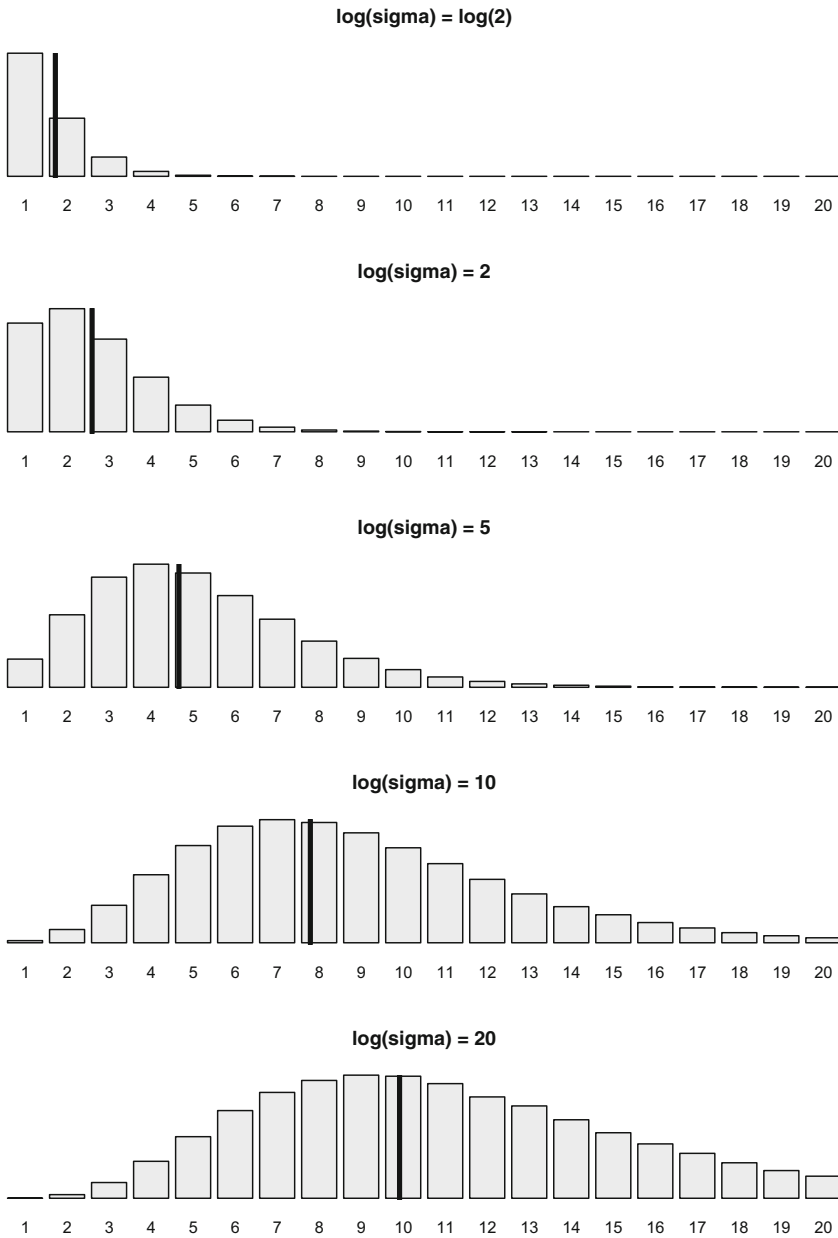


Fig. 7 Distributions of number of searches

## 6.4 Unobserved heterogeneity

The models estimated in this paper do not include unobserved heterogeneity (neither in preferences nor in search costs). However, we allow for observed heterogeneity

in search costs through the effects of distance to dealerships, an urban dummy, and an interaction between distance to dealership and the urban dummy. To identify unobserved heterogeneity in a continuous mixture model, a researcher must observe consumers more than once, i.e., the data must have a panel structure. In the case of search data, the researcher would have to observe multiple search spells per consumer (see, e.g., Dong et al. 2020).

In the following, we discuss several search papers that include unobserved heterogeneity in their models: Kim et al. (2010; 2017) include unobserved heterogeneity (in preferences and search costs) in their models and estimate them using aggregate data. While Yao et al. (2017) utilize both aggregate and individual-level data, in the estimation, the authors only include aggregate moments calculated using the individual-level data. Thus the interpretation of unobserved heterogeneity in these papers is the same as in Berry et al. (1995): unobserved heterogeneity is not distinguishable from a more flexible error structure. Chen and Yao (2017) and Honka (2014) both have cross-sectional, individual-level data. Chen and Yao (2017) include unobserved heterogeneity in both preferences and search costs, while Honka (2014) only includes unobserved heterogeneity in preferences. Identification of unobserved heterogeneity in these two papers comes from parametric assumptions.

Our data do not have a panel structure, i.e., we do not observe multiple search spells per consumer. In fact, given that we study the new car industry and the infrequent purchase occurrences in this market, it is unlikely that data on multiple search spells per consumer is/will be available. Even if such data were available, it is debatable whether the assumption of stable preferences would hold over such long time periods in an evolving market.

Therefore – because our data are cross-sectional, i.e., we observe one search spell per consumer – we include observed, but no unobserved heterogeneity in search costs in our model. If we were to include unobserved heterogeneity, identification of such would come from parametric assumptions.

A model with unobserved heterogeneity (especially in search costs; with MVSD fixed to 1) would also improve predictions in terms of the distribution of the number of searches (see Section 6.3) compared to a model with no unobserved heterogeneity (with MVSD fixed to 1). But because our data are cross-sectional, we only include observed heterogeneity in search costs in our model. If we had access to panel data on search, we could separately identify MVSD from search costs that also allow for unobserved heterogeneity.

## 7 Empirical results

We show the results from three search model specifications in Table 4. In model (i), we fix the MVSD to one and estimate a fixed cost of search. In model (ii), we continue to fix the MVSD to one, but allow search cost to vary with distance, an urban dummy, and an interaction between both variables. In model (iii), we allow the MVSD  $\sigma$  to enter the model as a parameter to be estimated. In all three model specifications, we include the same set of covariates as in the reduced-form choice models in Section 4. Across the three model specifications, the utility parameter estimates

**Table 4** Search model results

	Model (i)		Model (ii)		Model (iii)	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
<i>Preference parameters</i>						
MSRP (in \$10,000)	-0.182**	0.018	-0.186**	0.019	-0.202**	0.020
Chevrolet	-0.807**	0.035	-0.787**	0.034	-0.803**	0.038
Ford	-0.602**	0.037	-0.579**	0.039	-0.578**	0.040
Nissan	-0.068**	0.038	-0.063	0.040	-0.073	0.039
Toyota	0.047	0.031	0.072**	0.030	0.097**	0.033
Large vehicle x chevrolet	0.611**	0.059	0.611**	0.060	0.670**	0.066
Large vehicle x ford	0.886**	0.056	0.823**	0.060	1.020**	0.064
Large vehicle x toyota	0.028	0.053	0.011	0.054	0.046	0.058
Horsepower (in 100)	-0.473**	0.030	-0.501**	0.031	-0.518**	0.034
Engine size	0.132**	0.019	0.146**	0.018	0.151**	0.021
MPG	0.042**	0.005	0.042**	0.006	0.044**	0.006
MPG x large vehicle	0.005	0.003	0.006	0.003	0.004	0.004
<i>Log search cost parameters</i>						
Intercept	0.361**	0.018	0.273**	0.043	2.250**	0.040
Urban indicator			0.203**	0.048	-0.055**	0.015
Distance (in miles)			0.017**	0.001	0.006**	0.000
Urban x distance			0.002**	0.001	0.003**	0.000
<i>Match-value standard deviation</i>						
Sigma	1.00		1.00		8.16**	0.310
Number of consumers	6, 511		6, 511		6, 511	
Number of products	175,840		175,840		175,840	
Log-Likelihood	-20, 765		-18, 362		-17, 652	
BIC	41, 645		36, 872		35, 453	

Note: Asterisks indicate statistical significance at the 95% confidence level. The base brand is Honda. Optimization via BFGS with relative tolerance convergence criterion set to  $1e-6$ . Simulated likelihood using  $Q=1,000$  independent random draws of the random utility error and the match-value distribution. Search costs are parameterized as  $c_{ij} = \exp\{\gamma_0 + \gamma_1 \text{Urban} + \gamma_2 \text{Distance} + \gamma_3 \text{Urban} \times \text{Distance}\}$

are similar, generally sharing the same sign, similar magnitude, and significance with the exception of the brand intercepts for Nissan and Toyota. The utility parameter estimates also generally similar to the results from the reduced-form choice models presented in Table 2.

Comparing models (i) and (ii) in Table 4, the addition of distance, an urban dummy, and an interaction between both variables to the search cost function leads to a much larger log-likelihood value (a change of over 2,400). This improvement is driven by a better ability of the model to fit the search order during the estimation

process because consumers tend to visit dealerships close to their homes. When distance is not included in the model, two dealerships are equivalent from a modeling perspective if they offer a vehicle with the same characteristics even if one dealership is located next to the consumer's home and the other dealership is located 100 miles away.

In model (iii), our main model, we additionally also estimate the MVSD. The improvement of over 700 in the log-likelihood is large and shows that model (iii) fits the data better than model (ii). In addition, the correlation between  $\log(\sigma)$  and the search cost intercept  $\gamma_0$  is 0.95. If a correlation between two parameters equals  $\pm 1$ , then those two parameters are not separately identified. While a correlation of 0.95 is high, it is not high enough to suggest concern for identification.<sup>23</sup> The correlations between all other estimated coefficients lie within  $\pm 0.8$  with most being smaller than  $\pm 0.4$ .

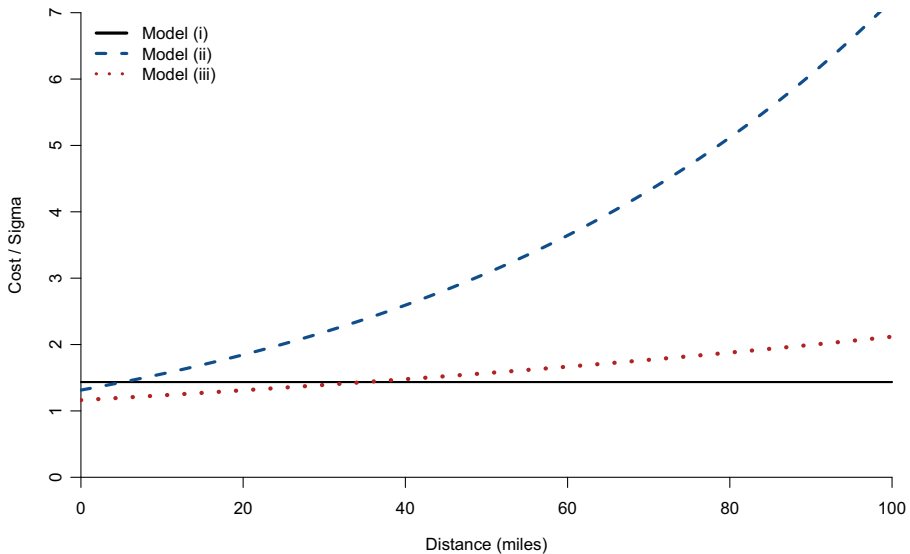
The MVSD estimate is 8.16. This estimate is large compared to the common practice of setting  $\sigma$  to 1. This large MVSD estimate indicates that consumers gain substantial benefits when visiting car dealerships. The MVSD has an estimated standard error of 0.31 and a 95% confidence interval (3.88, 17.19).<sup>24</sup> Note that the interval does not include 1. The parameter estimates for model (iii) also indicate that fixing the MVSD to 1 imposes a bias on the other parameters, in particular, on the cost parameters. More specifically, we find that the cost intercept is smaller, the urban intercept has a different sign, and the coefficient on distance is estimated to be almost three times as large in model (ii) as in model (iii). To demonstrate the differences visually, in Fig. 8, we plot  $c_{ij}/\sigma$  for a rural consumer over distances ranging from 0 to 100 miles for all three search models. The  $c_{ij}/\sigma$  value is substantially different between model (ii) and (iii) for almost all positive distances. Thus we conclude that it is important to estimate the MVSD to correctly recover search cost parameters.

It is important to note that the empirical findings related to the normalization of the MVSD to one, which we discuss in this and in the next section, are specific to our model and our empirical context. To put it differently, for a different product, the MVSD estimate might be close to one and thus normalizing it to one would be relatively innocuous. Our model also does not include unobserved heterogeneity – neither in preferences nor in search cost (see discussion in Section 6.4). It is an (empirical) open question what the consequences of normalizing the MVSD to one in a model with unobserved heterogeneity are. We speculate that (empirical) results will be biased, however, that the magnitude of the bias might be smaller. In that sense, comparisons between our empirical results based on a model without unobserved heterogeneity and empirical results from previous literature based on models with unobserved heterogeneity (e.g., Kim et al. 2010; Kim 2017; Chen and Yao 2017; Dong et al. 2020) should be conducted carefully.

Recall that our data are conditional on purchase. This does not affect the MVSD estimate since this parameter is identified by consumers' search behavior, i.e., search

<sup>23</sup>An alternative method to evaluate identification is to look at loglikelihood contour plots. These are available from the authors upon request and confirm the separate identification of  $\log(\sigma)$  and  $\gamma_0$ .

<sup>24</sup>We use the Delta method to calculate the standard error for  $\sigma$ .



**Fig. 8** Comparison of cost-sigma ratio vs distance across fitted search models (rural consumers)

order and number of searches. Thus MVSD is not affected by the inclusion of consumers who search, but do not end up making a purchase (as long as these consumers come from the same population). Our data are also conditional on search. For consumers who did not search (and did not buy a new car), searching must not have been optimal. Searching is optimal when the maximum reservation utility is larger than the utility of the outside option. Reservation utilities are a function of consumer preferences, search cost, and MVSD. In general, consumers are less likely to search if they have lower preferences, higher search cost, and a smaller MVSD. It is unclear whether and how MVSD would change if non-searchers were to be included in the data. This is the case because different scenarios are possible: MVSD might be the same, but non-searchers might have lower preferences or higher search costs. MVSD might be smaller making search less attractive. MVSD might be larger, but lower preferences and higher search cost might outweigh the effect of a larger MVSD. Thus we conclude that it is unclear how the MVSD estimate would change if non-searchers were to be included in the data.

Zooming in on search cost, we find urban consumers to have smaller search cost than rural consumers for dealerships fewer than 20 miles from their home. For larger distances, rural consumers' search cost are smaller than urban consumers' search cost. Following previous literature studying demand for cars and car search (e.g., Albuquerque and Bronnenberg 2012; Moraga-Gonzalez et al. 2015; Nurski and Verboven 2016; Murry and Zhou 2020), we also calculate distance-related search costs. For rural and urban consumers, our search cost estimates are \$297 and \$446 per mile, respectively. Compared to Albuquerque and Bronnenberg (2012) (\$64 per mile), Moraga-Gonzalez et al. (2015) (\$203 per mile), Nurski and Verboven (2016) (\$212 per mile), and Murry and Zhou (2020) (\$45 per mile), our estimates are somewhat higher. One potential explanation for this finding is that our data are more recent than

previous literature's. Over the last two decades, the number of dealership visits consumers conduct has decreased considerably. This decline might be partially driven by higher travel cost (Table 4).

## 7.1 Model fit

We evaluate the in-sample predictive performance of the estimated models using several measures. First, in Table 5, we show the product and brand hit rates for four models: full information multinomial logit model (model (b) in Table 2) and the three search models (models (i) to (iii) in Table 4). The “hit rate” is the percent of time that simulated behavior matches observed behavior. For example, the product hit rate is the average percent of time that the simulated-to-be-purchased product is the same as the actually purchased product.<sup>25</sup> The results show that all three search models considerably outperform the multinomial logit model (model (b) in Table 2). Among the three search models, added flexibility yields better hit rates. The product hit rate improves by about the same amount when search costs are modeled as a function of distance and the urban dummy (model (ii) in Table 4) and when additionally the MVSD is estimated (model (iii) in Table 4).

Additionally, in the bottom half of Table 5, we display the distribution of the number of searches consumers make (observed in our data) and the distributions of the number of searches that are predicted by the three search models from Table 4. While allowing search cost to vary with search cost shifters (model (ii)) enables the model to more precisely predict which products (dealerships) consumers search (see Section 7), model (ii) does not predict the distribution of the number of searches well: it overpredicts the proportion of consumers who search once and underpredicts the proportion of consumers who make two or more searches, i.e., the long tail of the distribution of the number of searches. While the maximum number of searches consumers conduct is 6+ in our data, model (ii) predicts that the maximum number of searches is three. Model (iii) – in which the MVSD is estimated – provides by far the best fit to the observed distribution of number of searches. In other words, permitting the MVSD to be freely estimated allows the model to fit the data better – in particular, the long tail of the distribution of the number of searches. Model (iii) predicts the maximum number of searches to be 5, while the maximum number of search in our data is 6+. To summarize, our results show that estimating the MVSD is crucial to correctly predict the distribution of the number of searches and especially its long tail.

## 7.2 Price elasticities

We show the implied own-price elasticities for all three search models in Table 6. Elasticities are calculated by first simulating search and choice behavior from the

---

<sup>25</sup>We calculate the hit rates as follows: for each consumer, we simulate 500 vectors of the random utility errors  $\eta_i$ . For each simulation, we calculate the consumer's reservation utilities for each searchable product and allow search and choice to progress following Weitzman's rules. For each consumer and each simulation, we record the predicted search length, the searched product(s), and the chosen product.



**Table 5** In-sample predictive performance

	Model (b)	Model (i)	Model (ii)	Model (iii)
<i>Hit Rates in %</i>	from Table 2	from Table 4	from Table 4	from Table 4
Product Hit Rate	8.4	10.4	14.9	19.4
Brand Hit Rate	30.0	39.4	39.8	39.9
	Observed	Model (i)	Model (ii)	Model (iii)
<i>Number of Searches</i>	(Data)	from Table 4	from Table 4	from Table 4
		(Predicted)	(Predicted)	(Predicted)
1	91.1%	96.2%	98.7%	88.5%
2	7.8%	3.7%	1.3%	10.3%
3	0.9%	0.1%	0.1%	1.1%
4	0.2%	–	–	0.1%
5	0.1%	–	–	0.1%
6+	0.1%	–	–	–

fitted model. Then, separately for each brand, we increase the prices of available alternatives by 10% and re-simulate search and choice behavior. We repeat this exercise 500 times for each consumer. Elasticities are computed as the average difference in percent between the simulated outcomes with and without a 1% price increase.

The own-price elasticity estimates range from -0.6 to -0.8. The elasticity estimates are influenced by the following two assumptions: first, our model is conditional on consumers having selected the type of vehicle they want to purchase. At some level of price increase or price decrease, consumers would choose a different type of vehicle for purchase. Our model (and thus the price elasticities) does not capture this behavior. And second, at some level of price increase or price decrease, consumers might want to search a brand that is not among the five brands under study. Our model (and thus the price elasticities) does not capture such behavior. While these two assumptions were useful in easing the computational burden of fitting the model, they restrict the interpretability of the elasticity estimates.

**Table 6** Own-price elasticities

	Model (i)	Model (ii)	Model (iii)
Chevrolet	-0.79	-0.78	-0.74
Ford	-0.70	-0.69	-0.67
Honda	-0.60	-0.59	-0.57
Nissan	-0.74	-0.74	-0.72
Toyota	-0.62	-0.61	-0.59

### 7.3 Consumer surplus

While the MVSD is a direct measure of the magnitude of potential benefits achievable through search, we also calculate consumer surplus as an alternative measure of benefits from search. For an individual consumer, consumer surplus is defined as the utility from the chosen product net of all costs (price and search costs), i.e.,

$$\mathbb{E}[CS_i] = \int_{\varepsilon} \int_{\eta} \left( u_{ij^*} - \sum_{j=1}^{K_i} c_{ij} \right) dF(\eta) dF(\varepsilon). \tag{12}$$

Note that, because our data are conditional on search and on purchase and thus we do not model the no-search and no-purchase decisions, our consumer surplus estimates taking both prices and search costs into account can be negative (see, e.g., Moraga-Gonzalez et al. 2017). To calculate consumer surplus for a consumer, we simulate 500 realizations from a model and average the resulting consumer surplus estimates for each consumer:

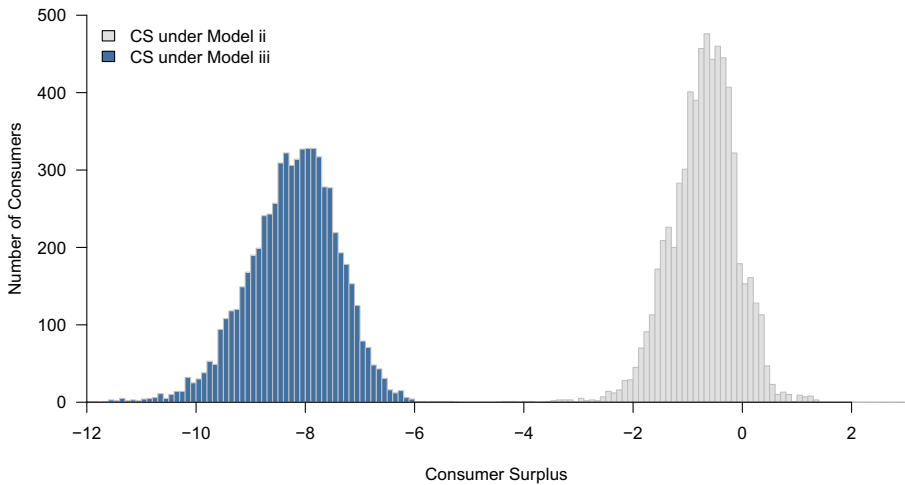
$$\widehat{CS}_i = \frac{1}{Q} \left( \sum_{q=1}^Q \left( u_{ij^*}^q - \sum_{j=1}^{K_i^q} c_{ij}^q \right) \right). \tag{13}$$

Figure 9 shows the distributions of consumer surplus from models (ii) and (iii). The mean consumer surplus estimates for models (ii) and (iii) are -0.7 and -8.0, respectively. Thus normalizing the MVSD to one severely *overstates* consumer surplus. Consumer surplus estimates from model (iii) are smaller than those from model (ii) for the following reason: the estimated MVSD in model (iii) is 8.16 and thus much larger than the value of one to which it is commonly fixed. To match up with observed search behavior, the search cost estimate in model (iii) is estimated to be much larger than the search cost estimate in model (ii). The consumer surplus formula shown in Equation (11) takes all cost into account – both price and search cost. Thus higher cost are subtracted from the utility of the chosen product and the consumer surplus is smaller.

Next, we assess the relationship between consumer surplus, consumer characteristics, and distance. To do so, we regress consumer surplus from model (iii), model (ii), and the difference in consumer surplus estimates between the two models on a set of demographics and distance.<sup>26</sup> The results are displayed in Table 7. Overall, demographics and distance explain a large proportion of variation in consumer surplus: 40% based on model (iii) and 34% based on model (ii). When it comes to the difference in consumer surplus between models (iii) and (ii), demographics and distance explain 58% of the variation in the difference in consumer surplus.

We find living in an urban area and belonging to a racial minority to be associated with higher consumer surplus. Being older, having more kids, and a larger

<sup>26</sup>Distance (for each consumer) is measured as the average distance across all searched dealerships in Table 7. In Appendix D, we assess the robustness of our results using two alternative measures of distance: average distance to all searchable dealerships (in Table D-1) and distance to selling dealership (in Table D-2). The results are similar.



**Fig. 9** Consumer surplus based on models (ii) and (iii)

distance to dealerships is associated with lower consumer surplus. We find no significant association between gender, income, education, and unemployment rate and consumer surplus. Our results are similar to Lee (2019) who finds that younger adults have higher consumer surplus from the adoption of smartphone devices than older adults. While our results show no significant association between consumer surplus and household income, Lee (2019) finds that households with higher incomes have higher consumer surplus. Lastly, four variables are significant in predicting the difference in consumer surplus between model (iii) and model (ii): age between 30 - 40 years (-), number of children (-), belonging to a racial minority (+), and distance (-).

To summarize, in the new car market, normalizing the MVSD to one results in a severe overestimation of consumer surplus. Demographics and distance explain a large proportion of variation in consumer surplus, namely, 40% in this market. And lastly, when it comes to buying a new car, consumer surplus is significantly higher for consumers belonging to a racial minority and living in an urban area, while it is significantly lower for older consumers and those with children.

## 8 Counterfactual

In 2017, Hyundai initiated a company-wide program called “Hyundai Drive” in which a consumer can schedule a test drive of a Hyundai vehicle at a location that is convenient to her and a dealer will bring the car to that location.<sup>27</sup> This program reduces travel costs to a dealer, but does not eliminate all search costs as it takes time

<sup>27</sup>Some dealers selling other-branded cars may also be offering at-home test drives. However, these are dealership-specific programs. To the best of our knowledge, Hyundai is the only car brand that has a company-wide program to offer at-home test drives.

**Table 7** Consumer surplus and demographics

	Model (iii)		Model (ii)		Model (iii) - Model (ii)	
	(1)	(2)	(3)	(4)	(5)	(6)
Urban indicator	0.792** (0.023)	0.666** (0.029)	-0.069** (0.019)	-0.230** (0.022)	0.861** (0.011)	0.896** (0.015)
Median age 30 - 40	-0.025 (0.024)	-0.076** (0.023)	0.035 (0.020)	-0.052** (0.018)	-0.060** (0.012)	-0.024** (0.012)
Median Age 40 +	-0.115** (0.030)	-0.141** (0.030)	-0.050** (0.025)	-1.116** (0.023)	-0.065** (0.015)	-0.025 (0.015)
Percent male	0.228 (0.178)	0.241 (0.170)	0.184 (0.149)	0.135 (0.131)	0.044 (0.089)	0.106 (0.088)
log(Number of kids)	-0.062** (0.010)	-0.072** (0.010)	-0.040** (0.008)	-0.056** (0.007)	-0.022** (0.005)	-0.015** (0.005)
Percent with college degree	0.215** (0.051)	-0.001 (0.051)	0.281** (0.043)	-0.006 (0.040)	-0.066** (0.026)	0.005 (0.027)
log(Income in \$1,000)	0.016** (0.007)	0.0002 (0.007)	0.020** (0.006)	0.002 (0.006)	-0.004 (0.004)	-0.002 (0.004)
Unemployment rate	0.238 (0.207)	0.097 (0.201)	0.302 (0.173)	0.115 (0.156)	-0.065 (0.104)	-0.019 (0.105)
Percent black	0.675** (0.069)	0.346** (0.072)	0.789** (0.058)	0.317** (0.055)	-0.114** (0.034)	0.029 (0.037)
Percent other race	0.946** (0.080)	0.644** (0.081)	0.972** (0.067)	0.517** (0.062)	-0.025 (0.040)	0.127** (0.042)
Distance	-0.010** (0.001)	-0.010** (0.001)	-0.002 (0.001)	-0.004** (0.001)	-0.008** (0.001)	-0.006** (0.001)
Intercept	-8.921** (0.129)	-8.178** (0.134)	-1.058** (0.108)	-0.039 (0.104)	-7.863** (0.064)	-8.139** (0.070)
Month fixed effects	No	Yes	No	Yes	No	Yes
County fixed effects	No	Yes	No	Yes	No	Yes
Number of observations	6,511	6,511	6,511	6,511	6,511	6,511
R <sup>2</sup>	0.284	0.393	0.080	0.337	0.539	0.578

Note: Standard errors reported in parentheses. Asterisks indicate statistical significance at the 95% confidence level. Distance is the average (across searches for one consumer) distance in miles to searched dealerships.

to take the test drive and the mental costs of considering an alternative continue to be present. Here, we assess the effects from adopting at-home test drive programs.

Our empirical results show that distance-related cost represent, on average, 6% (8%) of total search costs for rural (urban) consumers. When investigating the effects of a brand’s at-home test drive programs, we set the distance-related search costs to this brand’s closest dealership to zero. Thus, the consumer incurs no travel costs to “visit” that dealership. To calculate the impact of adopting at-home test drives, we

first simulate search and choice behavior for each consumer 500 times. Then, for each brand, we set the distance of the closest dealership for each consumer to zero and re-simulate search and choice behavior. We then calculate the average difference in search decisions and brand choices from the base simulation to the counterfactual simulation.

We first study the effects of unilateral at-home test drive program implementations for each of the five brands. We investigate changes in three outcomes: the probability that a consumer searches a brand, the probability that a consumer purchases a brand conditional on searching it, and market shares (unconditional purchase probabilities). The results are presented in the top half of Table 8. Unilateral implementation of at-home tests drives permits brands to capture an additional 3.9–7.7 percentage points of the analyzed market. Because the five brands under study account for 60.1% of the Texas auto market, this corresponds to overall market share changes of 2.3–4.6 percentage points (holding constant the choices made by consumers who purchase other brands). If we decompose the purchase decisions into search and conditional purchase decisions, we see, not surprisingly, that the primary effect is on consumer search.<sup>28</sup>

Comparing the market share increases in percent across the five car brands, we find that at-home test drive programs yield the smallest increases for the American car manufacturers Chevrolet and Ford and the largest increases for the Asian car manufacturers Honda, Nissan, and Toyota. This “asymmetric” effect is not driven by a change in preferences, but by the dealership structure: Chevrolet and Ford have about 210 dealerships each in Texas, while Honda, Nissan, and Toyota each have 55 to 80 dealerships (see Table 1). Thus, on average, consumers have to drive farther to get to a Honda, Nissan, or Toyota dealership than to a Chevrolet or Ford dealership, i.e., distance is a larger component of search cost for the former than the latter group of dealerships. With the implementation of at-home test drive programs, search cost to visit Honda, Nissan, and Toyota dealerships decrease more (in percent) than search cost to visit Chevrolet and Ford dealerships making the former relatively more attractive to search than the latter. The relatively larger increase in the searches for Honda, Nissan, and Toyota than for Chevrolet and Ford carries through to the purchase decisions increasing the market shares (in percent) of the former more than of the latter.

As a point of comparison, we conduct the same exercises with the estimates from model (ii) in which the MVSD is fixed to one. We find the results to be roughly one third to one fourth of those from model (iii) in which the MVSD is estimated. Specifically, estimated increases in market shares under model (ii) range from 0.8 to 1.6 percentage points. Thus, failing to estimate MVSD yields underestimated market share increases from the unilateral implementation of at-home test drives.

Next, we investigate the effects of three more scenarios: when Toyota and Honda (close substitutes) implement at-home test drive programs, when Toyota and Ford (weak substitutes) implement at-home test drive programs, and when all five brands

<sup>28</sup> A decrease in search costs makes consumer search more, i.e., increases the size of consumers’ consideration sets. As consideration sets increase in size, conditional purchase probabilities decrease because the denominator increases.

**Table 8** Unilateral adoption of at-home test drives

Brand adopting At-Home test drives	No At-Home Test drives	With At-Home Test drives	Change in Percent
<i>Model (iii)</i>			
Market share			
Honda	0.145	0.194	34.2%
Chevrolet	0.200	0.247	23.4%
Ford	0.286	0.338	18.0%
Nissan	0.093	0.132	41.3%
Toyota	0.276	0.353	28.0%
Probability of search			
Honda	0.163	0.221	35.4%
Chevrolet	0.225	0.281	24.4%
Ford	0.323	0.384	18.8%
Nissan	0.106	0.151	42.7%
Toyota	0.311	0.402	29.0%
Probability of purchase conditional on search			
Honda	0.904	0.896	-0.9%
Chevrolet	0.903	0.898	-0.6%
Ford	0.911	0.907	-0.5%
Nissan	0.897	0.886	-1.2%
Toyota	0.909	0.903	-0.7%
<i>Model (ii)</i>			
Market share			
Honda	0.148	0.164	11.3%
Chevrolet	0.198	0.211	6.4%
Ford	0.283	0.298	5.1%
Nissan	0.096	0.109	13.8%
Toyota	0.275	0.302	9.8%
Probability of search			
Honda	0.149	0.167	11.9%
Chevrolet	0.201	0.214	6.8%
Ford	0.287	0.303	5.5%
Nissan	0.097	0.111	14.4%
Toyota	0.279	0.307	10.3%
Probability of purchase conditional on search			
Honda	0.990	0.985	-0.5%
Chevrolet	0.988	0.985	-0.3%
Ford	0.989	0.987	-0.3%
Nissan	0.988	0.982	-0.6%
Toyota	0.990	0.986	-0.4%

implements at-home test drive programs. The results are shown in Table 9. In the two scenarios in which two brands implement at-home test drive programs, all brands implementing such a program increase their market shares. While Toyota and Honda approximately equally benefit from at-home test drive programs, Toyota increases its market shares about three times more than Ford. This result is consistent with the brands' number of dealerships in Texas. When all five brands implement at-home test drive programs, Honda, Nissan, and Toyota increase their market shares, while Chevrolet and Ford lose market shares. Not surprisingly, the more brands implement at-home test drive programs, the smaller are the average market share changes (in percent). In percentage points (after accounting for the five brands under study having 60.1% of the Texas auto market), Honda, Nissan, and Toyota increase their market shares by 0.66–0.72 percentage points and Chevrolet and Ford decrease their market shares by 0.72–1.38 percentage points. Thus at-home test drive programs are a more effective tool for car manufacturers with a thin dealership network.

To summarize, we find that the primary effect of at-home test drive programs is, not surprisingly, on consumer search. The effect on conditional purchase is minor. Failure to estimate the MVSD results in an underprediction of market share changes due to at-home test drive programs by a factor 3 to 4. Further, brands with fewer dealerships benefit more from at-home test drive programs than brands with more dealerships. This is the case because consumers have to travel farther to get to those dealerships. Thus at-home test drive programs result in a bigger search cost reduction for brands with fewer dealerships, making these brands much more attractive to search. In Texas, brands like Ford and Chevrolet have about three to four times more dealerships than brands like Honda, Nissan, and Toyota. Thus the three Asian car manufacturers in our data benefit more from at-home test drive programs than the two U.S. car manufacturers. This “asymmetric” effect on market shares is not driven by a change in consumer preferences, but by a change in search costs.

## 9 Limitations and future research

Our paper is not without limitations and offers opportunities for future research. First, for searched but not purchased vehicles, we observe that a consumer visited a dealership, but not which specific car the consumer was interested in. This is a limitation of our data. We therefore make the assumption that the consumer was searching for similar cars across different dealerships. Second and related to the first point, we do not observe whether a consumer considered one or multiple vehicles at a dealership. We therefore make the assumption that the consumer considered one vehicle. This assumption is supported by industry reports which suggest that consumers have a specific car type in mind when visiting dealerships.<sup>29</sup> Third, we do not model vehicle

<sup>29</sup>See, e.g.:

<https://www.edmunds.com/car-buying/10-steps-to-finding-the-right-car-for-you.html>;

<https://santanderconsumerusa.com/learning-center/i-need-another-car-now-what>;

[https://growwithcars.com/wp-content/uploads/2014/05/2014.05-Consumer-Journey-Report\\_Final.pdf](https://growwithcars.com/wp-content/uploads/2014/05/2014.05-Consumer-Journey-Report_Final.pdf).



**Table 9** Additional counterfactual predictions based on model (iii)

Brand	No At-Home Test drives	With At-Home Test drives	Change in Percent
<i>Toyota &amp; Honda</i>			
Market share			
Honda	0.145	0.172	19.2%
Chevrolet	0.200	0.172	-14.0%
Ford	0.286	0.247	-13.7%
Nissan	0.093	0.081	-13.8%
Toyota	0.276	0.328	19.0%
Probability of search			
Honda	0.163	0.197	20.4%
Chevrolet	0.225	0.194	-14.1%
Ford	0.323	0.279	-13.6%
Nissan	0.106	0.091	-13.7%
Toyota	0.311	0.374	20.0%
Probability of purchase conditional on search			
Honda	0.904	0.894	-1.1%
Chevrolet	0.903	0.901	-0.3%
Ford	0.911	0.908	-0.3%
Nissan	0.897	0.895	-0.3%
Toyota	0.909	0.901	-0.9%
<i>Toyota &amp; Ford</i>			
Market share			
Honda	0.145	0.121	-16.2%
Chevrolet	0.200	0.163	-18.8%
Ford	0.286	0.305	6.6%
Nissan	0.093	0.079	-16.0%
Toyota	0.276	0.333	20.7%
Probability of search			
Honda	0.163	0.137	-16.1%
Chevrolet	0.225	0.183	-18.7%
Ford	0.323	0.347	7.4%
Nissan	0.106	0.089	-15.9%
Toyota	0.311	0.379	21.7%
Probability of purchase conditional on search			
Honda	0.904	0.902	-0.2%
Chevrolet	0.903	0.901	-0.3%
Ford	0.911	0.904	-0.8%
Nissan	0.897	0.894	-0.3%
Toyota	0.909	0.901	-0.8%

**Table 9** (continued)

Brand	No At-Home Test drives	With At-Home Test drives	Change in Percent
<i>All brands</i>			
Market share			
Honda	0.145	0.156	8.1%
Chevrolet	0.200	0.188	-5.9%
Ford	0.286	0.263	-8.1%
Nissan	0.093	0.104	11.5%
Toyota	0.276	0.288	4.6%
Probability of search			
Honda	0.163	0.179	9.3%
Chevrolet	0.225	0.214	-5.1%
Ford	0.323	0.299	-7.3%
Nissan	0.106	0.119	12.8%
Toyota	0.311	0.329	5.6%
Probability of purchase conditional on search			
Honda	0.904	0.892	-1.4%
Chevrolet	0.903	0.893	-1.2%
Ford	0.911	0.900	-1.3%
Nissan	0.897	0.883	-1.5%
Toyota	0.909	0.897	-1.3%

choice, only brand-dealer choice. It is left for future research to extend the model to also incorporate vehicle choice.

Fourth, as common in the search literature, we assume that the match value is independent of the part of utility that is observed by the consumer prior to searching. This assumption allows us to use the three decision rules developed by Weitzman (1979). At the same time, it is an assumption that limits the flexibility of our model. In recent work, Gardete and Anthill (2020) relax this assumption and allow for correlation between the match value and the part of utility that is observed by the consumer prior to searching. And lastly, the benefit provided by dealers might be heterogenous varying with brand, dealership size or vehicle class. We leave it for future research to explore these types of heterogeneity.

## 10 Conclusion

Prior to the Internet, store visits and the physical inspection of products played a prominent role during consumers' purchase process – especially for high-ticket durable goods such as automobiles. Information on prices and vehicle features was difficult to obtain except via a dealership visit. During the last 15 years, a wealth of

online information sources have provided a low cost alternative to time consuming dealership visits. Therefore it is an important empirical question whether dealerships continue to provide (substantial) value to consumers in the Internet age.

To answer this question, it is necessary to separately quantify both the cost *and* the benefit of search. We estimate a sequential search model for product fit and show that – with an exogenous search cost shifter – both the cost and the benefit of search can be estimated. Our empirical results show that the benefit provided by dealerships to consumers remains substantial. Consumers value visiting dealerships and learning about their liking (or dislike) of more experiential product attributes. This finding points to a continuously important role of physical stores in the Internet age. The actions of online retailers in other product categories also corroborate this conclusion. Initially online-only retailers, such as Warby Parker and Bonobos, have recently opened brick and mortar stores to provide consumers with information that is difficult to communicate electronically.

Using our empirical results, we calculate consumer surplus and relate it to consumer demographics. We find that a surprisingly large proportion of the variation in consumer surplus in the new car market can be explained by demographics. And lastly, we evaluate the effects of at-home test drive programs which decrease consumers' search cost. We observe that at-home test drive programs increase market shares of car brands with thin dealership networks and decrease market shares of car brands with thick dealership networks, i.e., have asymmetric effects. How changes in search costs affect market structure for different products and in different markets warrants further examination especially due to the recent search cost increases in offline shopping, i.e., store visits, due to COVID-19.

## **Appendix A: Details on Analysis Data Set Construction**

### **Combining Search and Purchase Data**

There is no unique identifier in the Safegraph dealership visit data that corresponds to information in the Texas DMV registrations data to indicate which search sequence corresponds to which purchased vehicle. Therefore, to merge these two datasets, we employ the following algorithm: first, the algorithm finds a search sequence that includes a visit to the selling dealership prior to the vehicle registration date, and second, the algorithm requires that the home location of the searcher and of the buyer are approximately the same. In the following, we provide more details on the employed algorithm.

First, recall that the home location of a mobile device user in the search data is estimated by Safegraph and provided as a geohash-8 while the purchase data from the Texas DMV provides the registrant's street address. We geocode both sets of location information as latitude and longitude coordinates. For the search data, we use the center of the geohash; for the purchase data, we use coordinates from the GoogleMaps API.

We then require that the two home locations are in close proximity. To do so, we calculate the distance between each of the 154,000 mobile device home locations in

the dealership visits data and each of the 264,000 unique registrant's home addresses in the purchase data using the Haversine great-circle distance measure, which is the shortest distance between two points on a sphere. Due to computer memory constraints, for each mobile device home location, we retain up to a maximum of 50 "potential merges" (i.e., registered vehicles) where the distance between the home locations is one-fifth of a mile or closer. In short, we find up to 50 vehicle sales that could potentially be the result of each observed search sequence.

Next, we evaluate potential merges in increasing order of distance. For each potential merge, we assess whether the mobile device had visited the selling dealership and whether that visit occurred on or before the vehicle registration date. If so, we merge that search sequence with that purchased vehicle, and we discard all other potential merges involving that search sequence or vehicle. If, instead, the potential merge involved home locations that were close in proximity but where the mobile device had not visited the selling dealership on or before the registration date, we discard that potential merge. We then evaluate the next (in terms of shortest distance) potential merge. The process proceeds until all potential merges are evaluated and either accepted or discarded.

With this algorithm, we find the purchased vehicle for 12,065 search sequences. This represents 9.5% of the mobile devices in the Safegraph data and 6.2% of the vehicle registrations in the Texas DMV data. These figures appear reasonable for the following reasons: first, the search data also include dealership visits for consumers who were shopping for used vehicles or who were taking a currently owned vehicle to a dealership's service department for maintenance – both activities are likely to occur with greater frequency than new car shopping. Second, the search data also include dealership visits by consumers who searched, but did not purchase a new vehicle. And third, the search data purportedly represent 5%–10% of U.S. mobile devices and thus it is expected that these data capture approximately the same percentage of new car purchases.

### The Searchable Set of Dealerships

In much the same way that a multinomial discrete choice model requires information on the chosen product as well as the products not chosen, our structural search model requires information on the set of (potentially) searchable products. Across the 35 brands, there are more than 1,100 dealerships in the state of Texas. To ease the computational burden imposed by such a large set of searchable products, we take two steps.

First, we limit the analysis to the five brands with the largest market shares in Texas during the sample period, which are Chevrolet, Ford, Honda, Nissan, and Toyota. This set of brands has a combined market share of 60%. Focusing the analysis on these brands requires that we also limit the data to consumers who purchased these brands and whose search histories only include visits to dealerships that carry these brands. This reduces the sample size to 7,051 consumers.

Second, we limit the searchable set of dealerships to any dealership within a particular radius of the consumer – twice the 70th percentile of the distance between

the home location of a car buyer and the selling dealership. We do so by county and also differentiating between rural and urban consumers. In the following, we provide more details on this data cleaning procedure.

We limit the searchable set of dealerships to any dealership within a particular radius of the consumer. To define the radius for each consumer, we first partition consumers into urban and rural consumers. Urban consumers are those who live within the city limits of any major city (where major is defined as a city with a population exceeding 100,000 inhabitants) and rural consumers live outside of these major cities. For urban consumers in each major city and for rural consumers in each county (excluding inhabitants of the major cities in that county), we assign a radius as twice the 70th percentile of the distance between the home location of a car buyer and the selling dealership. We calculate these distance distributions from the original data set containing 195,000 Texas DMV new vehicle registrations.

For example, suppose a consumer lives in the rural part of Dallas County and our data contain 10,000 new vehicle registrations with the Texas DMV from rural residents of Dallas County. We calculate the 70th percentile of the distance between those 10,000 residents' home and the selling dealerships. Suppose this value is 30 miles. Then for each consumer in the analysis sample who lives in the rural part of Dallas County, we assign a radius of 60 miles. Therefore, the searchable set of dealerships for rural Dallas County residents includes any dealership within 60 miles of the consumer's home location. Note that this set of dealerships may differ across rural residents of Dallas County if some dealerships are within 60 miles of some consumers' homes, but others are not.

The average radiuses for urban and rural consumers are 29.6 and 33.5 miles, respectively. We find that 7.6% of consumers search or purchase a vehicle from a dealership outside of their assigned radius and exclude these consumers from the analysis sample. This step reduces the sample size to 6,511 consumers making 7,175 dealership visits.

## **The Searchable Set of Vehicles**

Through the process of merging the search and purchase data, we know which vehicle each individual consumer purchased at which selling dealership. However, for all dealerships she visited but did not make a purchase, we do not observe which vehicle she searched. Further, for all dealerships the consumer decided not to visit, we do not observe which vehicle she would have searched if she had decided otherwise. To estimate the structural search model, we need both pieces of information. We make the following set of assumptions about searched and potentially searchable vehicles.

For each consumer and each searchable dealership identified by the process outlined in the previous section, we assume that there is one most preferred vehicle that could be searched, i.e., if the consumer were to visit a particular dealership, she would have searched only her most preferred vehicle at that dealership. It is important to note that the most preferred vehicle at a particular dealership is consumer-specific: if the same dealership is searchable by multiple consumers, each consumer has her own most preferred vehicle at that dealership. To identify this consumer-specific, most

preferred vehicle at each dealership, we impose two criteria: the potentially searched vehicle must be in that dealership's inventory and it must be of the same Edmunds type as the vehicle ultimately purchased by the consumer. The latter assumption is equivalent to assuming that consumers search for a specific vehicle conditional on having decided the type of vehicle they would like to buy.

To identify the consumer-specific, most preferred vehicle at each dealership, we impose two criteria: the potentially searched vehicle must be in that dealership's inventory and it must be of the same Edmunds type as the vehicle ultimately purchased by the consumer. More specifically, "inventory" is defined as any model sold by that dealership during the 16 months for which the Texas DMV registration data are available.<sup>30</sup> This criteria enforces that consumers can only search for a model at a particular dealership if that dealership has ever carried and sold that model. In addition to being in inventory, the searched vehicle must be of the same Edmunds type as the vehicle ultimately purchased by the consumer. This assumption is equivalent to assuming that consumers search for a specific vehicle conditional on having decided the type of vehicle they would like to buy.

To continue the example, suppose the rural Dallas County consumer is observed to purchase a 2017 Nissan Versa Note and that her searchable set of dealerships includes XYZ Chevrolet. According to Edmunds, two Chevrolet vehicles are of the same type (type: "Extra-Small Hatchback") as the Nissan Versa Note: the Chevrolet Spark and the Chevrolet Sonic. Suppose that the Texas DMV registrations data show that the XYZ Chevrolet dealership sold zero 2017 Chevrolet Sparks and nineteen 2017 Chevrolet Sonics. Then the searchable vehicle for our example consumer at XYZ Chevrolet is assumed to be the 2017 Chevrolet Sonic because (i) it has the same model year and the same Edmunds type as the purchased vehicle and because (ii) it was in inventory at the searched dealership.<sup>31</sup>

Finally, we must also infer the set of vehicle characteristics for each searchable vehicle. Following previous literature (e.g., Albuquerque and Bronnenberg 2012, Moreno and Terwiesch 2017), we impute these characteristics. To do so, we compute the median values (for numeric characteristics) and modal values (for categorical characteristics) by model *and* dealership. Thus if the rural Dallas County consumer could search a 2017 Chevrolet Sonic at two different Chevrolet dealers (say ABC Chevrolet and XYZ Chevrolet), the city mileage of the 2017 Sonic at ABC Chevrolet depends on the city mileage of the other 2017 Sonics sold by ABC Chevrolet and differs from the city mileage calculated for the 2017 Sonic at XYZ Chevrolet.

## Appendix B: Distance to Dealerships by Education, Race, and Age

<sup>30</sup> A model is a particular combination of model year, make, and model. For example, one model is a 2017 Toyota Corolla.

<sup>31</sup> In the event that multiple vehicles meet these criteria, the vehicle that is ranked highest according to Edmunds.com is selected as the most-preferred vehicle.

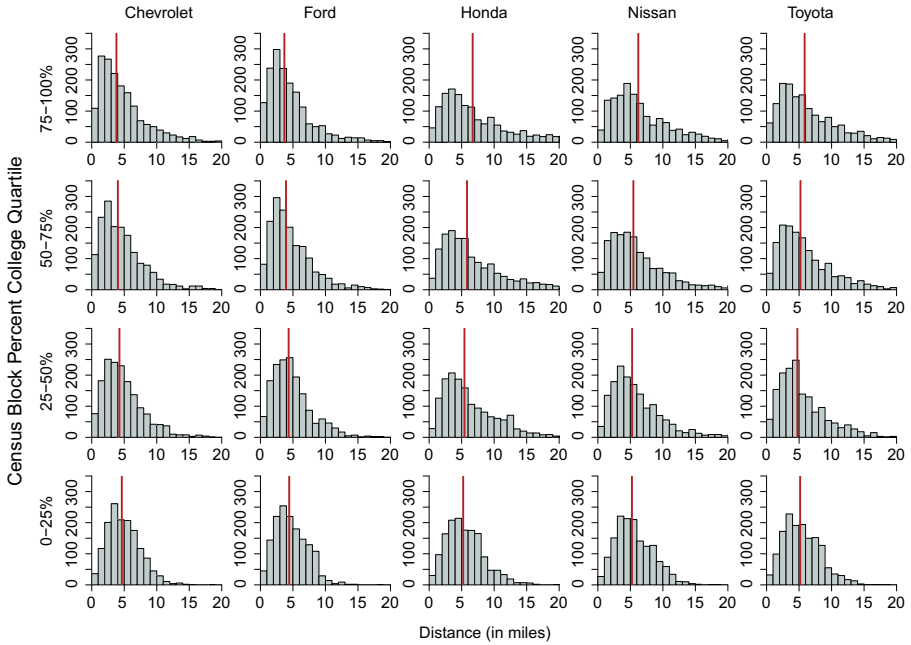


Fig. 10 Distance histograms by brand and education

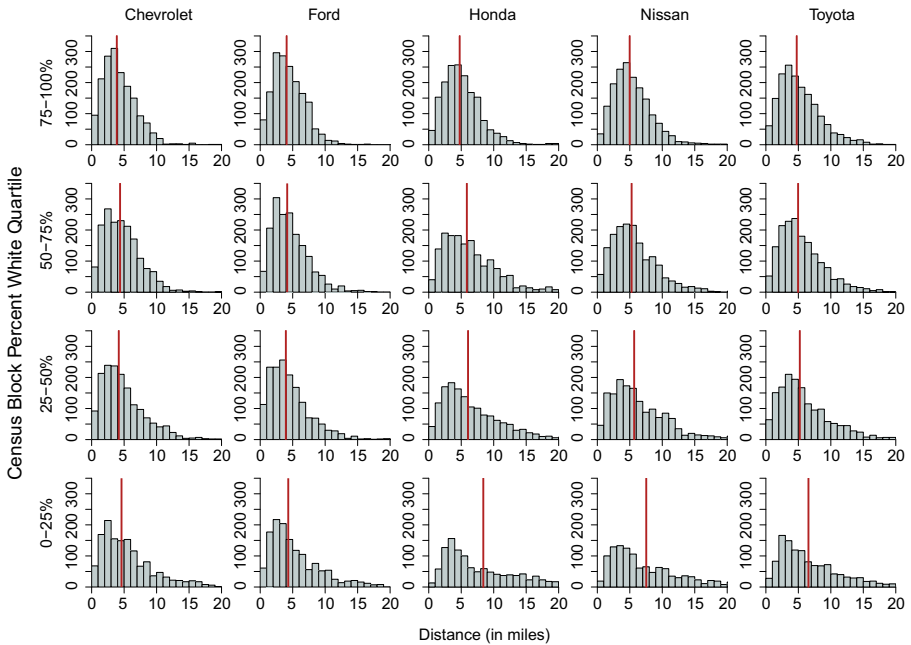


Fig. 11 Distance histograms by brand and race

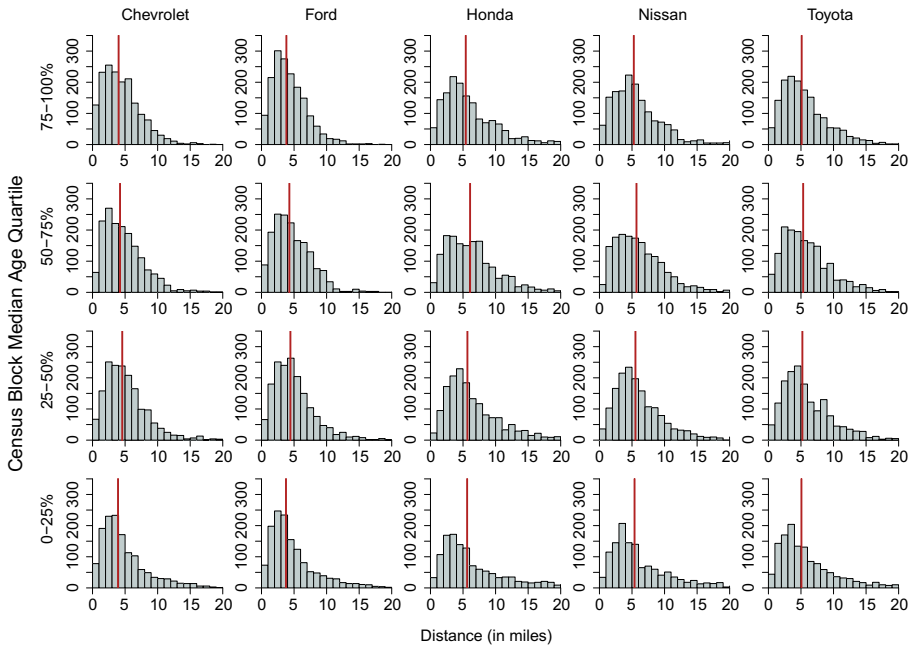


Fig. 12 Distance histograms by brand and age

## Appendix C: Kernel-Smoothed Frequency Simulator

Neither the search nor purchase probabilities can be expressed in closed form. We approximate the integrals in the likelihood function with averages using logit-smoothed accept-reject simulation. This simulated maximum likelihood estimation algorithm follows Train (2009) and is outlined in the following steps:

1. Do steps 2–4 for each consumer  $i = 1, \dots, N$
2. Draw  $Q$  values from the density  $f_\varepsilon(\varepsilon)$  for each of the  $K_i$  searches for a total of  $Q \cdot K_i$  draws and draw  $Q$  values from the density  $f_\eta(\eta)$  for a total of  $Q \cdot J_i$  draws
3. For each set of random draws:
  - (a) Compute  $v_{1,j} = z_{ij} - \max_{h \leq j} \{u_{ih}\}$  for  $j = 2, \dots, K_i$
  - (b) Compute  $v_{2,j} = z_{ij} - \max_{k > j} \{z_{ik}\}$  for  $j = 1, \dots, K_i$
  - (c) Compute  $v_3 = \max_{h \leq K_i} \{u_{ih}\} - \max_{k > j} \{z_{ik}\}$
  - (d) Compute  $v_4 = u_{ij_i^*} - \max_{h \leq K_i} \{u_{ih}\}$  for the chosen  $j_i^*$
  - (e) Compute the simulated individual likelihood given one set of draws:

$$\tilde{L}_i^q = \left( 1 + \sum_{j=2}^{K_i} e^{-\lambda_1 v_{1,j}} + \sum_{j=1}^{K_i} e^{-\lambda_2 v_{2,j}} + e^{-\lambda_3 v_3} + e^{-\lambda_4 v_4} \right)^{-1}$$



4. Average the draw-specific simulated individual likelihoods over the draws:

$$\tilde{L}_i = \frac{1}{Q} \sum_{q=1}^Q \tilde{L}_i^q$$

5. Take logs and aggregate the individual-specific likelihoods over individuals to form the total simulated log-likelihood function:

$$\log(\tilde{L}) = \sum_{i=1}^N \log(\tilde{L}_i)$$

Computing time when fitting the sequential search model is directly proportional to the number of draws used to approximate the integrals in the likelihood function. However, a large number of draws is necessary to achieve a good approximation. When calculating each individual likelihood, we therefore use 1,000 draws for  $\eta_{ij}$  and 1,000 draws for  $\varepsilon_{ij}$ . By comparison, Kim et al. (2017) use 40 draws when testing the kernel-smoothed AR estimation approach; Honka (2014) and Ursu (2018) use 50 draws to estimate their search models with the same approach.

### Appendix D: Consumer Surplus Regressions

**Table 10** Consumer surplus and demographics: distance to searchable dealerships

	Model (iii)		Model (ii)		Model (iii) - Model (ii)	
	(1)	(2)	(3)	(4)	(5)	(6)
Urban indicator	0.791** (0.023)	0.664** (0.029)	-0.071** (0.019)	-0.232** (0.022)	0.861** (0.011)	0.896** (0.015)
Median age 30 - 40	-0.023 (0.024)	-0.075** (0.023)	0.035 (0.020)	-0.052** (0.018)	-0.059** (0.012)	-0.024 (0.015)
Median age 40 +	-0.113** (0.030)	-0.139** (0.030)	-0.049** (0.025)	-0.115** (0.023)	-0.063** (0.015)	-0.024 (0.015)
Percent male	0.229 (0.178)	0.243 (0.170)	0.185 (0.149)	0.136 (0.131)	0.044 (0.089)	0.106 (0.088)
log(Number of Kids)	-0.062** (0.010)	-0.072** (0.010)	-0.040** (0.008)	-0.056** (0.007)	-0.022** (0.005)	-0.015** (0.005)
Percent with college degree	0.216** (0.051)	0.001 (0.051)	0.281** (0.043)	-0.006 (0.040)	-0.065** (0.026)	0.006 (0.027)
log(Income in \$1,000)	0.016** (0.007)	-0.001 (0.007)	0.020** (0.006)	0.002 (0.006)	-0.004 (0.004)	-0.002 (0.004)

**Table 10** (continued)

	Model (iii)		Model (ii)		Model (iii) - Model (ii)	
	(1)	(2)	(3)	(4)	(5)	(6)
Unemployment rate	0.238 (0.207)	0.098 (0.201)	0.302 (0.173)	0.115 (0.155)	-0.064 (0.104)	-0.017 (0.105)
Percent black	0.674** (0.069)	0.344** (0.072)	0.788** (0.058)	0.317** (0.055)	-0.114** (0.034)	0.027 (0.037)
Percent other race	0.946** (0.080)	0.644** (0.081)	0.971** (0.067)	0.516** (0.062)	-0.025 (0.040)	0.128** (0.042)
Distance	-0.010** (0.001)	-0.011** (0.001)	-0.002 (0.001)	-0.004** (0.001)	-0.008** (0.001)	-0.006** (0.001)
Intercept	-8.921** (0.129)	-8.176** (0.134)	-1.057** (0.108)	-0.037 (0.104)	-7.864** (0.064)	-8.139** (0.579)
Month fixed effects	No	Yes	No	Yes	No	Yes
County fixed effects	No	Yes	No	Yes	No	Yes
Number of observations	6,511	6,511	6,511	6,511	6,511	6,511
R <sup>2</sup>	0.285	0.394	0.081	0.337	0.540	0.584

Note: Standard errors reported in parentheses. Asterisks indicate statistical significance at the 95% confidence level. Distance is the average distance in miles to searchable dealerships

**Table 11** Consumer surplus and demographics: distance to selling dealership

	Model (iii)		Model (ii)		Model (iii) - Model (ii)	
	(1)	(2)	(3)	(4)	(5)	(6)
Urban indicator	0.792** (0.023)	0.666** (0.029)	-0.069** (0.019)	-0.230** (0.022)	0.861** (0.011)	0.896** (0.015)
Median age 30 - 40	-0.025 (0.024)	-0.076** (0.023)	0.035 (0.020)	-0.052** (0.018)	-0.060** (0.012)	-0.024** (0.012)
Median age 40 +	-0.115** (0.030)	-0.141** (0.030)	-0.050** (0.025)	-1.116** (0.023)	-0.065** (0.015)	-0.025 (0.015)
Percent male	0.228 (0.178)	0.241 (0.170)	0.184 (0.149)	0.135 (0.131)	0.044 (0.089)	0.106 (0.088)
log(Number of Kids)	-0.062** (0.010)	-0.072** (0.010)	-0.040** (0.008)	-0.056** (0.007)	-0.022** (0.005)	-0.015** (0.005)
Percent with college degree	0.215** (0.051)	-0.001 (0.051)	0.281** (0.043)	-0.006 (0.040)	-0.066** (0.026)	0.005 (0.027)
log(Income in \$1,000)	0.016** (0.007)	0.0002 (0.007)	0.020** (0.006)	0.002 (0.006)	-0.004 (0.004)	-0.002 (0.004)
Unemployment rate	0.238	0.097	0.302	0.115	-0.065	-0.019

**Table 11** (continued)

	Model (iii)		Model (ii)		Model (iii) - Model (ii)	
	(1)	(2)	(3)	(4)	(5)	(6)
	(0.207)	(0.201)	(0.173)	(0.156)	(0.104)	(0.105)
Percent black	0.675**	0.346**	0.789**	0.317**	-0.114**	0.029
	(0.069)	(0.072)	(0.058)	(0.055)	(0.034)	(0.037)
Percent other race	0.946**	0.644**	0.972**	0.517**	-0.025	0.127**
	(0.080)	(0.081)	(0.067)	(0.062)	(0.040)	(0.042)
Distance	-0.010**	-0.010**	-0.002	-0.004**	-0.008**	-0.006**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Intercept	-8.921**	-8.178**	-1.058**	-0.039	-7.863**	-8.139**
	(0.129)	(0.134)	(0.108)	(0.104)	(0.064)	(0.070)
Month fixed effects	No	Yes	No	Yes	No	Yes
County fixed effects	No	Yes	No	Yes	No	Yes
Number of observations	6,511	6,511	6,511	6,511	6,511	6,511
R <sup>2</sup>	0.284	0.393	0.080	0.337	0.539	0.578

Note: Standard errors reported in parentheses. Asterisks indicate statistical significance at the 95% confidence level. Distance is the distance in miles to the selling dealership

## Appendix E: Complete Counterfactual Predictions

**Table 12** Toyota and honda adopt at-home test drives

Brand	No At-Home Test Drives	With At-Home Test drives	Change in Percent
<i>Model (iii)</i>			
Market share			
Honda	0.145	0.172	19.2%
Chevrolet	0.200	0.172	-14.0%
Ford	0.286	0.247	-13.7%
Nissan	0.093	0.081	-13.8%
Toyota	0.276	0.328	19.0%
Probability of search			
Honda	0.163	0.197	20.4%
Chevrolet	0.225	0.194	-14.1%
Ford	0.323	0.279	-13.6%
Nissan	0.106	0.091	-13.7%
Toyota	0.311	0.374	20.0%

**Table 12** (continued)

Brand	No At-Home Test Drives	With At-Home Test drives	Change in Percent
Probability of purchase conditional on search			
Honda	0.904	0.894	-1.1%
Chevrolet	0.903	0.901	-0.3%
Ford	0.911	0.908	-0.3%
Nissan	0.897	0.895	-0.3%
Toyota	0.909	0.901	-0.9%
<i>Model (ii)</i>			
Market share			
Honda	0.148	0.157	6.6%
Chevrolet	0.198	0.188	-5.2%
Ford	0.283	0.269	-5.0%
Nissan	0.096	0.092	-4.5%
Toyota	0.275	0.294	6.9%
Probability of search			
Honda	0.149	0.160	7.2%
Chevrolet	0.201	0.190	-5.2%
Ford	0.287	0.273	-5.0%
Nissan	0.097	0.093	-4.5%
Toyota	0.279	0.299	7.4%
Probability of purchase conditional on search			
Honda	0.990	0.985	-0.5%
Chevrolet	0.998	0.988	0.0%
Ford	0.990	0.989	0.0%
Nissan	0.988	0.988	0.0%
Toyota	0.990	0.986	-0.4%

**Table 13** Toyota and ford adopt at-home test drives

Brand	No At-Home Test drives	With At-Home Test drives	Change in Percent
<i>Model (iii)</i>			
Market share			
Honda	0.145	0.121	-16.2%
Chevrolet	0.200	0.163	-18.8%
Ford	0.286	0.305	6.6%
Nissan	0.093	0.079	-16.0%
Toyota	0.276	0.333	20.7%
Probability of search			
Honda	0.163	0.137	-16.1%
Chevrolet	0.225	0.183	-18.7%
Ford	0.323	0.347	7.4%
Nissan	0.106	0.089	-15.9%
Toyota	0.311	0.379	21.7%
Probability of purchase conditional on search			
Honda	0.904	0.902	-0.2%
Chevrolet	0.903	0.901	-0.3%
Ford	0.911	0.904	-0.8%
Nissan	0.897	0.894	-0.3%
Toyota	0.909	0.901	-0.8%
<i>Model (ii)</i>			
Market share			
Honda	0.148	0.140	-5.3%
Chevrolet	0.198	0.185	-6.5%
Ford	0.283	0.287	1.2%
Nissan	0.096	0.091	-4.9%
Toyota	0.275	0.297	8.0%
Probability of search			
Honda	0.149	0.141	-5.3%
Chevrolet	0.201	0.188	-6.5%
Ford	0.287	0.292	1.6%
Nissan	0.097	0.092	-4.9%
Toyota	0.279	0.302	8.4%
Probability of purchase conditional on search			
Honda	0.990	0.990	0.0%
Chevrolet	0.998	0.988	0.0%
Ford	0.990	0.987	-0.3%
Nissan	0.988	0.988	0.0%
Toyota	0.990	0.986	-0.4%

**Table 14** All brands simultaneously adopt at-home test drives

Brand	No At-Home Test drives	With At-Home Test drives	Change in Percent
<i>Model (iii)</i>			
Market share			
Honda	0.145	0.156	8.1%
Chevrolet	0.200	0.188	-5.9%
Ford	0.286	0.263	-8.1%
Nissan	0.093	0.104	11.5%
Toyota	0.276	0.288	4.6%
Probability of search			
Honda	0.163	0.179	9.3%
Chevrolet	0.225	0.214	-5.1%
Ford	0.323	0.299	-7.3%
Nissan	0.106	0.119	12.8%
Toyota	0.311	0.329	5.6%
Probability of purchase conditional on search			
Honda	0.904	0.892	-1.4%
Chevrolet	0.903	0.893	-1.1%
Ford	0.911	0.900	-1.2%
Nissan	0.897	0.883	-1.4%
Toyota	0.909	0.897	-1.1%
<i>Model (ii)</i>			
Market share			
Honda	0.148	0.153	3.4%
Chevrolet	0.198	0.191	-3.3%
Ford	0.283	0.272	-3.9%
Nissan	0.096	0.101	5.3%
Toyota	0.275	0.283	2.7%
Probability of search			
Honda	0.149	0.155	3.9%
Chevrolet	0.201	0.195	-2.9%
Ford	0.287	0.277	-3.5%
Nissan	0.097	0.103	5.9%
Toyota	0.279	0.288	3.2%
Probability of purchase conditional on search			
Honda	0.990	0.985	0.5%
Chevrolet	0.988	0.984	0.4%
Ford	0.990	0.986	0.3%
Nissan	0.988	0.981	0.6%
Toyota	0.990	0.985	0.5%

## References

- Albuquerque, P., & Bronnenberg, B.J. (2012). Measuring the impact of negative demand shocks on car dealer networks. *Marketing Science*, *31*(1), 4–23.
- Berry, S., Levinsohn, J., Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, *63*(4), 841–890.
- Bucklin, R.E., Siddarth, S., Silva-Risso, J.M. (2008). Distribution intensity and new car choice. *Journal of Marketing Research*, *45*(4), 473–486.
- Chen, Y., & Yao, S. (2017). Sequential search with refinement: Model and application with Click-Stream data. *Management Science*, *63*(12), 4345–4365.
- De los Santos, B., & Koulayev, S. (2017). Optimizing click-through in online rankings with endogenous search refinement. *Marketing Science*, *36*(4), 542–564.
- Dong, X., Morozov, I., Seiler, S., Hou, L. (2020). Estimation of Preference Heterogeneity in Markets with Costly Search. Working paper.
- Gardete, P., & Anthill, M. (2020). Guiding Consumers through Lemons and Peaches: An Analysis of the Effects of Search Design Activities. Working Paper.
- Honka, E. (2014). Quantifying search and switching costs in the U.S. Auto insurance industry. *The RAND Journal of Economics*, *45*(4), 847–884.
- Honka, E., & Chintagunta, P. (2017). Simultaneous or sequential? search strategies in the U.S. Auto insurance industry. *Marketing Science*, *36*(1), 21–42.
- Hui, S., Bradlow, E., Fader, P. (2009). Testing behavioral hypotheses using an integrated model of grocery store shopping path and purchase behavior. *Journal of Consumer Research*, *36*(3), 478–493.
- Hui, S., Inman, J., Huang, Y., Suher, J. (2013). The effect of In-Store travel distance on unplanned spending: Applications to mobile promotion strategies. *Journal of Marketing*, *77*(2), 1–16.
- Jain, A., Misra, S., Rudi, N. (2016). Sales Assistance, Search and Purchase Decisions: An Analysis Using Retail Video Data, Working Paper.
- Kim, J.B., Albuquerque, P., Bronnenberg, B.J. (2010). Online demand under limited consumer search. *Marketing Science*, *29*(6), 1001–1023.
- Kim, J.B., Albuquerque, P., Bronnenberg, B.J. (2017). The probit choice model under sequential search with an application to online retailing. *Management Science*, *63*(11), 3911–3929.
- Koulayev, S. (2014). Search for differentiated products: Identification and estimation. *The RAND Journal of Economics*, *45*(3), 553–575.
- Lee, S. (2019). Quantifying the Effects of Smartphone Adoption: Digital Device Substitution and Digital Consumption Expansion, Working Paper.
- McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, *57*(5), 995–1026.
- Moraga-Gonzalez, J., Sandor, Z., Wildenbeest, M. (2015). Consumer Search and Prices in the Automobile Market. Working paper.
- Moraga-Gonzalez, J., Sandor, Z., Wildenbeest, M. (2018). Consumer Search and Prices in the Automobile Market. Working paper.
- Moraga-Gonzalez, J.L., Sandor, Z., Wildenbeest, M.R. (2017). Nonsequential search equilibrium with search cost heterogeneity. *International Journal of Industrial Organization*, *50*, 392–414.
- Moreno, A., & Terwiesch, C. (2017). The effects of product line breadth: Evidence from the automotive industry. *Marketing Science*, *36*(2), 254–271.
- Morozov, I. (2019). Measuring Benefits from New Products in Markets with Information Frictions. Working paper.
- Morton, F.S., Silva-Risso, J., Zettelmeyer, F. (2011). What matters in a price negotiation: Evidence from the US auto retailing industry. *Quantitative Marketing and Economics*, *9*(4), 365–402.
- Murry, C., & Zhou, Y. (2020). Consumer search and automobile dealer Co-Location. *Management Science*, *66*(5), 1909–1934.
- Nurski, L., & Verboven, F. (2016). Exclusive dealing as a barrier to entry? evidence from automobiles. *The Review of Economic Studies*, *83*(3), 1156–1188.
- Palazzolo, M., & Feinberg, F. (2015). Modeling Consideration Set Substitution. Working paper.
- Ratchford, B.T., & Srinivasan, N. (1993). An empirical investigation of returns to search. *Marketing Science*, *12*(1), 73–87.
- Ratchford, B.T., Talukdar, D., Lee, M.-S. (2007). The impact of the internet on consumers' use of information sources for automobiles: A Re-Inquiry. *Journal of Consumer Research*, *34*(1), 111–119.

- Seiler, S., & Pinna, F. (2017). Estimating search benefits from Path-Tracking data: Measurement and determinants. *Marketing Science*, 36(4), 565–589.
- Seiler, S., & Yao, S. (2017). The impact of advertising along the conversion funnel. *Quantitative Marketing and Economics*, 15(3), 241–278.
- Train, K.E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press, 2nd edn.
- Ursu, R. (2018). The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Science*, 37(4), 530–552.
- Ursu, R., Wang, Q., Chintagunta, P. (2020). Search Duration. *Marketing Science*, 39(5), 849–1031.
- Weitzman, M.L. (1979). Optimal search for the best alternative. *Econometrica*, 47(3), 641–654.
- Yao, S., Wang, W., Chen, Y. (2017). TV Channel search and commercial breaks. *Journal of Marketing Research*, 54(5), 671–686.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Affiliations

Dan Yavorsky<sup>1</sup> · Elisabeth Honka<sup>1</sup>  · Keith Chen<sup>1</sup>

Elisabeth Honka  
elisabeth.honka@anderson.ucla.edu

Keith Chen  
keith.chen@anderson.ucla.edu

<sup>1</sup> University of California, Los Angeles, Los Angeles, CA, USA