INSEAD
The Business School for the World®

Working Paper Series
2018/36/TOM

Reading between the Stars: Understanding the Effects of Online Customer Reviews on Product Demand

Hallie S. Cho
INSEAD, hallie.cho@insead.edu

Sameer Hasija
INSEAD, sameer.hasija@insead.edu

Manuel E. Sosa
INSEAD, manuel.sosa@insead.edu

Problem definition: Many studies have examined quantitative customer reviews (i.e., star ratings) and found them to be a reliable source of information that has a positive effect on product demand. Yet the effect of qualitative customer reviews (i.e., text reviews) on demand has been less thoroughly studied, and it is not known whether (or how) the sentiment expressed in text reviews moderates the influence of star ratings on product demand. We are therefore led to examine how the interplay between review sentiment and star ratings affects product demand. *Academic/practical relevance*: Consumer perceptions of product quality - and how they are shared via customer reviews - are of extreme relevance to the firm, but we still do not understand how product demand is affected by the quantitative and qualitative aspects of customer reviews. Our paper seeks to fill this critical gap in the literature by analyzing star ratings, the sentiment of customer reviews, and their interaction. *Methodology*: Using 2002-2013 data for the US automobile market, we investigate empirically the impact of star ratings and review sentiment on product demand. Thus we estimate an aggregated multinomial choice model after performing a machine learning-based sentiment analysis on the entire corpus of customer reviews included in our sample. *Results*: We find that (i) review sentiment and star ratings both have a decreasingly positive effect on product demand and (ii) the effect (on demand) of their interaction suggests that the two components of reviews are complements. Positive sentiments in text reviews compensate for the tendency of consumers to discount extremely high star ratings, and negative sentiments amplify that discounting tendency. *Managerial implications*: The firm should pay greater attention to quantitative and qualitative customer reviews so that it can better understand how consumers perceive the quality of its products or services.

Keywords: Perceived Quality; Product Demand; Online Customer Reviews; Sentiment Analysis

Acknowledgement:
The authors thank Xu Chi (scientist at SIMTech) and A*STAR for their assistance with some portions of the sentiment analysis featured in this paper.

Reading between the Stars: Understanding the Effects of Online Customer Reviews on Product Demand

Hallie S. Cho
INSEAD, hallie.cho@insead.edu,

Sameer Hasija
INSEAD, sameer.hasija@insead.edu,

Manuel E. Sosa
INSEAD, manuel.sosa@insead.edu,

Problem definition: Many studies have examined quantitative customer reviews (i.e., star ratings) and found them to be a reliable source of information that has a positive effect on product demand. Yet the effect of qualitative customer reviews (i.e., text reviews) on demand has been less thoroughly studied, and it is not known whether (or how) the sentiment expressed in text reviews moderates the influence of star ratings on product demand. We are therefore led to examine how the interplay between review sentiment and star ratings affects product demand. *Academic/practical relevance*: Consumer perceptions of product quality—and how they are shared via customer reviews—are of extreme relevance to the firm, but we still do not understand how product demand is affected by the quantitative and qualitative aspects of customer reviews. Our paper seeks to fill this critical gap in the literature by analyzing star ratings, the sentiment of customer reviews, and their interaction. *Methodology*: Using 2002–2013 data for the US automobile market, we investigate empirically the impact of star ratings and review sentiment on product demand. Thus we estimate an aggregated multinomial choice model after performing a machine learning–based sentiment analysis on the entire corpus of customer reviews included in our sample. *Results*: We find that (i) review sentiment and star ratings both have a decreasingly positive effect on product demand and (ii) the effect (on demand) of their interaction suggests that the two components of reviews are complements. Positive sentiments in text reviews compensate for the tendency of consumers to discount extremely high star ratings, and negative sentiments amplify that discounting tendency. *Managerial implications*: The firm should pay greater attention to quantitative *and* qualitative customer reviews so that it can better understand how consumers perceive the quality of its products or services.

Key words: perceived quality, product demand, online customer reviews, sentiment analysis

1. INTRODUCTION

When considering the purchase of a product whose quality is either unknowable ex ante or difficult to assess, consumers tend to rely on signals by looking to others for that information (Festinger 1954, Deutsch and Gerard 1955). In the days of traditional media, these sources of information were friends, neighbors, colleagues, and endorsements from celebrities or experts. With the rise

of social media and the proliferation of online customer reviews, consumers can consult others on a much larger scale, thereby taking advantage of the wisdom generated by crowds of product reviewers (Avery et al. 1999, Dellarocas 2003, 2006). Product reviewers share their opinion about the quality of a product after experiencing it, which we refer to as *ex post quality*.

Recent research in operations management acknowledges the increasing presence of online product review platforms and have studied their impact on product design strategies, pricing, firm profitability, and consumer welfare (Yu et al. 2016, Crapis et al. 2017, Feldman et al. 2018). But just how do these online review platforms influence product demand? On such platforms, product reviewers who have experienced the product provide a quantitative rating—a "star rating"—and/or a qualitative assessment in terms of textual product reviews. Although several studies have shown that product reviews influence product demand (Godes and Mayzlin 2004, Chevalier and Mayzlin 2006, Liu 2006, Dellarocas et al. 2007, Duan et al. 2008, Forman et al. 2008, Ghose et al. 2012, Lehman et al. 2014, Luca 2016), we still do not understand how quantitative and qualitative ratings *interact* to affect product demand. The goal of our paper is to shed some light on this important question.

Previous studies that relate product reviews to product demand focus mainly on the effects of quantitative ratings. This approach makes sense given that star ratings are not only readily available but also perhaps the most salient feature of product reviews (Dellarocas et al. 2007, Luca 2016). Salience is crucial because of an interesting trade-off in harnessing wisdom from crowds. A greater number of reviews ensures that the crowd's average estimate is closer to the product's true quality, but it also entails that the cost of consuming this information is increasing. It is a simple matter to average star ratings; however, the nature of text reviews is such that they cannot be averaged and so require the reader's time and effort to process all the text (Godes and Mayzline 2004). Accordingly, Chetty et al. (2009) showed that consumers overlook less salient information when purchasing consumables but that this generalization may not apply to big-ticket items. Yet Luca and Smith (2013) reported that consumers do rely on relatively coarse information even when contemplating such major expenditures as college education.

Despite being less salient and requiring more effort to process, text reviews are not entirely ignored by consumers. Some studies have shown that certain characteristics of text reviews directly affect product sales. Thus, according to Chevalier and Mayzlin (2006), the length of text reviews has a significantly positive effect on sales. Ghose and Ipeirotis (2011) controlled for average star ratings and still found that the readability of text reviews—as well as the text's ratio of subjective to objective statements—positively affects sales. The implications of these findings are twofold: first, text reviews are likely to contain information that is not captured by star ratings; second,

consumers may well read text reviews in order to assess the validity of the corresponding aggregated star ratings.

Scholars examining the text of product reviews have reported evidence of such comments providing specific information that is not captured in quantitative ratings. Bickart and Schindler (2001) found that text reviews tend to be more subjective yet also more credible than information provided by the manufacturer, since reviews describe specific usage scenarios and contexts. Netzer et al. (2012) used the co-mentions of *other* products within a review to create a brand association network and derive insights into the competitive landscape from the perspective of consumers. Archak et al. (2011) extracted mentions of product features and associated evaluative words ("bad", "good", "excellent", etc.) and found that the occurrence of such evaluations affects product demand. Thus most prior work in this area has examined either certain aspects or selected snippets of text reviews. In contrast, we evaluate the entire corpus of text reviews in our sample in order to estimate the overall "sentiment" that captures how reviewers feel about a given product. For this purpose we exploit advances in natural language processing and machine learning—an approach that allows us to create, from the text of reviews, a quantitative measure that is comparable in that respect to star ratings.

Some studies have treated ratings and sentiment as de facto substitutes by claiming that the former are a fair approximation of the latter (Chevalier and Mayzlin 2006, Ghose and Ipeirotis 2011). As Ghose and Ipeirotis stated, "the numerical rating score already gives the (approximate) polarity of the review" (2011, p. 1501). Indeed, one might reasonably suppose that a product's star rating and the sentiment of its text review both reflect the reviewer's opinion of that product and hence should be *substitutable* proxies of product quality. However, we propose that star ratings and review sentiment have instead a *complementary* effect on demand.

In particular, we draw on dual-process theory in psychology[1] to argue that star ratings are the result of "automatic" System 1 thinking, which is characterized as fast, intuitive, and emotional, whereas text reviews result from "effortful" System 2 thinking, which is characterized as slow, deliberative, and logical (Tversky and Kahneman 1974, Gilbert 1991, Epstein 1994, Sloman 1996, Stanovich and West 2000, Kahneman and Frederick 2002, Evans and Stanovich 2013). This conceptual model is consistent with how star ratings and review comments are typically solicited and used, and we therefore posit that text reviews play a moderating role in the effect of star ratings on product demand.

Toward the end of understanding how the interplay between quantitative and qualitative reviews affects product demand, we first examine how the aggregated ratings and the average sentiment of

[1] This concept of two types of thinking was popularized by Kahneman's (2011) book entitled *Thinking, Fast and Slow*.

text reviews actually relate to product demand. We argue (and find empirically) that star ratings and review sentiment are both positively related to product demand but that the effects are not linear: in general, product reviews positively affect product demand at a decreasing rate. We then study how the effects of ratings and sentiment interact and thereby assess the extent to which text reviews and star ratings are complements, as hypothesized. We find robust empirical evidence that the sentiment of product reviews positively moderates the effect of star ratings. It is interesting that, given the nonlinearity of the baseline effects of star ratings and sentiment, a positive interaction effect indicates that if star ratings are extremely positive then a positive sentiment *counteracts* the negative tendency to discount those ratings—whereas a negative sentiment *reinforces* that tendency to discount extremely positive star ratings.

We begin with a presentation of our theoretical framework, which shows how product reviews influence product demand; we then argue how the effect of star ratings on product demand is moderated by the sentiment of product reviews. Next, we evaluate our hypotheses empirically using longitudinal data on the US auto market for the period 2002–2013. This evaluation is based on our text analysis of automobile reviews, which was conducted via a supervised machine learning algorithm. Finally, we discuss how our findings contribute to the operations management literature on quality and its relationship to consumer purchasing decisions; we also draw managerial implications for the manufacturer and for hosts of product reviews.

2. THEORY AND HYPOTHESES

2.1. Experience attributes and the wisdom of crowds

The economics literature broadly categorizes a product as a "search good" or an "experience good" based on the difficulty of obtaining quality information prior to purchase (Feldman et al. 2018). The notion of search versus experience goods is from Nelson (1974): "goods can be classified by whether the quality variation was ascertained predominantly by search or by experience." For search goods, quality can be evaluated prior to purchase; for experience goods, however, quality cannot be evaluated without sampling or actually purchasing. Yet changes in product trends and consumers' information-seeking behavior have increased the number of products that do not fit neatly into Nelson's classification of search and experience goods.

With the distinction between a product and a service becoming blurred, there are more products whose quality can be determined only through a *combination* of search and experience. Especially for new technology products, assessing quality from product descriptions and specifications can be confusing and difficult. For products that are complex, determining their relative overall quality via an attribute-by-attribute comparison is a decidedly nontrivial task—even though information about these attributes can be easily obtained by searching (Ulrich and Ellison 1999, Terwiesch and Loch 2004, Randall et al. 2007).

One of the main principles of social psychology is that individuals look to others in order to reduce uncertainty in their own judgments (Festinger 1954). It is therefore hardly surprising that, when evaluating complex products whose quality is difficult to ascertain, consumers are likely to look to others (Feldman et al. 2018). Online product reviews offer consumers not only a means of viewing the opinions of many others but also information on users' evaluation of the focal product's ex post quality. We define ex post quality as the quality realized *after* the use or consumption of a product, which means that it can be determined only from experience. Ex post quality can be a good representation of the benefits that consumers can expect to gain from purchasing the product, but the evaluation process is highly subjective and prone to individual biases.

Because there are an extremely large number of reviews available online, concerns about subjectivity of the evaluations can be assuaged by leveraging the wisdom of crowds. Using the wisdom of crowds to assess a complex product's quality is based on the idea that, given a large enough crowd whose individual errors are independent and identically distributed (i.i.d.) and also unbiased, a simple average of the crowd's estimates should yield the best estimate of ex post product quality (Clemen and Winkler 1986, Clemen 1989, Graham 1996, Armstrong 2001). In the context of reviews, the individual biases that cast doubts on the accuracy of any individual's assessment can be canceled out by aggregating reviews. It follows that aggregate reviews provide the best estimate of a product's ex post quality and hence of what an average user can expect to experience after purchasing that product.

2.2. Ex post quality and product demand

Using aggregate reviews as the best estimation of ex post product quality is consistent with the literature on the wisdom of crowds, but consumer behavior does not actually reflect this way of thinking. Larrick and Soll (2006) found that people tend to believe that averaging estimates will lead to only average accuracy. Since individuals thus incorrectly view an averaged estimate as being more uncertain than it really is, it follows that—with respect to a product's ex post quality—the expected value of their utility is *less* than the utility of their expected value. These biases are likely to introduce nonlinearities when assessing the effect of product reviews on product demand.

Only a few studies have explored possible nonlinearities associated with the usefulness of a product's rating that are due to perceived uncertainty of aggregated product reviews. Mudambi and Schuff (2010) documented a nonlinear relationship between ratings and review "helpfulness"; extreme ratings are associated with low levels of helpfulness—but only for experience goods, not for search goods. These authors suggest that extreme reviews reduce helpfulness only for experience goods owing to the subjective nature of an experience. In addition, Maslowska et al. (2017) found support for an inverted-U relationship between ratings and purchase propensity; they argue that

there may be a "too good to be true" effect at play. Maslowska et al. reported that demand for a product peaks when its ratings are between 4.2 and 4.5 (out of 5) and that higher values are associated with *reduced* demand. In line with Mudambi and Schuff, they suspect that this dynamic reflects concerns about the credibility of extremely positive aggregated reviews.

We therefore expect the expected utility of ex post quality—as captured by product reviews—to be concave; that is, we suppose that marginal utility diminishes as ex post quality increases. Because ex post quality is communicated via product reviews, which are shared in quantitative and qualitative ways, we anticipate finding a nonlinear relationship between aggregated product reviews and product demand. More specifically, we expect that both aggregated star ratings and aggregated text sentiment have a positive but diminishing effect on product demand. We accordingly offer the following formal hypotheses.

Hypothesis 1a *(H1a). Aggregated star ratings have a positive but diminishing effect on product demand.*

Hypothesis 1b *(H1b). Aggregated sentiments of text reviews have a positive but diminishing effect on product demand.*

It is important to distinguish explicitly between the effects of aggregate star ratings and sentiment; this approach is consistent with the notion that they are quasi-independent signals of ex post quality—in other words, they are not simply reflections of one another. Having made this distinction, we can further explore the relationship between these two types of signals and how they influence a consumer's decision-making process.

2.3. Dual-process theory

The structure of online product reviews provides a unique opportunity to gain insights into how consumers perceive product quality and share those perceptions of quality. Most product review platforms request that reviewers give both star ratings and review text—and usually in that order (Luca 2016). So when reviewing a product, customers are first asked to rate it; this process captures the reviewer's "gut feeling" about the focal product. Once a star rating has been assigned, the reviewer is asked to comment further; this aspect of the review gives the customer an opportunity to reflect on and rationalize her decision by offering opinions about the product as well as accounts of her experience. We argue (i) that star ratings and textual comments are outputs from two *different* cognitive processes of the reviewer and (ii) that, analogously, star ratings and textual comments trigger those particular dictinct cognitive processes in potential consumers who consult such reviews.

Philosophers and psychologists have long maintained that there are two different types of reasoning, which are referred to as System 1 and System 2 thinking (Tversky and Kahneman 1974,

Gilbert 1991, Epstein 1994, Sloman 1996, Stanovich and West 2000, Kahneman and Frederick 2002, Evans and Stanovich 2013). System 1 thinking is characterized as being automatic, associative, experienced based, biased, and contextualized; in contrast, System 2 thinking is controlled, rule based, normative, abstract, and focused on the consequences of decision making. According to most scholars, the main differences are that System 1 thinking is fast and intuitive whereas System 2 thinking is slow and analytical. The "correction" model of Gilbert (1991) posited that System 2, a slower process that engages in hypothetical reasoning, can correct for the biases and errors typical of the judgments hastily made via System 1. Sloman (1996) added that System 1's automatic nature—the individual who employs this type of thinking is aware of the result but not of *how* it was arrived at—implies that the reasoning inherent in System 2 can suppress System 1 but not completely override it.

What is unique about the structure of product reviews is that they separately capture the reviewer's System 1 and System 2 responses. Reviewers are first asked to rate the product by giving it a certain number of stars, which requires only a single click and is likely to capture the reviewer's System 1 response. The reviewer is then given a text box in which write the product review, which is likely to capture his System 2 response. Star ratings are similar to an automatic response in this sense: the reviewer often cannot fully explain his logic for assigning the rating (System 1 thinking). When the reviewer is writing a review, however, he is exerting effort and using his working memory to analyze his experience and assess the product's quality (System 2 thinking). Because they capture both star ratings and text reviews from each individual reviewer, product reviews allow us to examine how a product's quality is perceived by these two different cognitive systems.

It is important not only to understand the cognitive processes associated with submitting a product review but also to consider the cognitive processes of the potential customers who consult product reviews. From the perspective of "digesting" product reviews, aggregated star ratings help form an intuitive first impression of the product whereas reading text reviews is likely to shape one's overall impression of the product. On product review platforms, the average numerical rating is displayed first (and most prominently); this rating is then followed by a series of individual text reviews. The aggregation process for text differs from that for rating. Although ratings can be numerically averaged, text comments are aggregated only in a sense of being displayed as a list that is accessible en masse—that is, they seldom undergo any further processing. Hence consumers must exert effort to process System 2 responses (text reviews) but not System 1 responses (star ratings).

Thus the reviewer's process of reviewing a product parallels a potential customer's process of consuming that review. Star ratings capture the reviewer's quickly formed (automatic) opinion

of the product, and a reader can digest that information just as quickly. Text reviews reflect the reviewer's effortful analysis of a product, and a reader must likewise exert effort to process all the information contained in that analysis.

Whereas a product's aggregated star rating will likely trigger a potential customer's System 1 reaction of forming a quick judgment about the product's ex post quality, her System 2 reasoning will likely be activated when she reads the text reviews. Then, the aggregated sentiment expressed in those reviews may confirm or moderate the initial assessment triggered by the aggregated star rating. We propose that there exists a positive synergy between ratings and sentiment similar to the recursive cycle described by Mandler (1982) and Rindova and Petkova (2007)—in which positive emotions increase comprehension, which in turn increases positive emotions. That is, additional positive emotion is generated if the newly learned information is also positive and thus endorses the initial reaction; however, contradictory new information induces skepticism and negative emotion. Hence we expect there to be a positive synergy between the star ratings and text comments of a product review when the latter are in tune with the former.

Given such a synergistic effect between aggregated star ratings and aggregated text sentiments, we anticipate that the marginal effect of star ratings on demand will be increasing in text sentiment. Recalling the decreasing return on ratings posited by Hypotheses 1, it follows that the positive complementary effect just described should mitigate the discounting of highly positive ratings. We express that expectation formally as follows.

Hypothesis 2 *(H2). The effect of aggregated star ratings on product demand is increasing in product reviews' aggregated sentiment. Hence, increasing product reviews' sentiment mitigate the tendency to discount extremely positive aggregated star ratings.*

3. ECONOMETRIC ANALYSIS

The context for our study is the US automotive industry. We examine how the ex post quality evaluations captured in product reviews influence the purchasing decisions of consumers in the market for a new car—a complex good whose quality can be ascertained through search and observed through experience. In order to test our hypotheses, we build a sample from several sources of data about new cars sold in the United States during the period 2002–2013. Our sample includes a car's search characteristics, experience characteristics, price, and other relevant factors that could affect a consumer's purchasing decision. Before giving further details on the data and variables, we shall discuss our econometric model and identification approach.

3.1. Econometric Model

We observe aggregate demand but are interested in understanding factors that influence an individual consumer's choice; therefore, we consider an aggregated form of the multinomial logit demand

model (McFadden 1973, 1986). This approach is consistent with consumer choice models used extensively in the literature on assortment planning (for a review, see Kök et al. (2015)). Hence we start by assuming that the consumer's utility is a function of product characteristics that can be learned via search or experience. Searchable product characteristics are objective measures that enter utility linearly, whereas experience characteristics are subjective measures that enter utility nonlinearly. We define the utility that a consumer i derives from purchasing a new car model j in year t as:

$$\mu_{ijt} = -\alpha \ln(\text{Price}_{jt}) + \sum_k \beta_k(\text{Search Characteristics}_{kjt}) + \sum_l \gamma_l(\text{Experience Characteristics}_{ljt}) - \sum_l \gamma_l'(\text{Experiene Characteristic}_{ljt}^2) + \sum_m \lambda_m(\text{Control Variables}_m) + \xi_{jt} + \epsilon_{ijt} \quad (1)$$

Search characteristics are price, horsepower, size, and miles per dollar; the term β_k is a vector of coefficients for these characteristics. *Experience* characteristics are star ratings and text sentiment, for which γ_l and γ_l' represent vectors of coefficients; λ_m is a vector of coefficients for control variables. Product characteristics that are unobservable to the econometrician are represented by ξ_{jt} , and ϵ_{ijt} is a mean-zero stochastic term that captures individual deviation from the population mean.

We assume a quadratic relationship between experience characteristics and utility. This quadratic utility structure is often used to model satiation (Economides et al. 2008, Kim et al. 2010). In our context, the expected negative effect of the quadratic experience characteristics may correspond to the "too good to be true" effect found by Maslowska et al. (2017)—that is, once we control for factors that may affect the information signal's credibility. Note also that the quadratic utility function is concave and can be monotonically increasing as long as the satiation point falls outside the range of interest.

Equation 1 can be simplified to read $\mu_{ijt} = V_{jt} + \epsilon_{ijt}$. We make the following assumptions: consumers are utility maximizers; consumer heterogeneity is captured by ϵ_{ijt} , which is i.i.d. and also Type I extreme value distributed; and the mean utility of outside goods (V_{0t}) can be normalized to 0. Hence we can find the individual choice probabilities by following the steps outlined in Train (1986, chap. 3) and then aggregate those probabilities to estimate market shares s_{jt} :

$$s_{jt} = \frac{e^{V_{jt}}}{1 + \sum_{k=1}^{K} e^{V_{kt}}}$$

Since the mean utility of outside goods is normalized to 0, the market share of outside goods can be written as $s_{0t} = 1/(1 + \sum_{k=1}^{K} e^{V_{kt}})$. Then the market share of car model j in year t simplifies to $s_{jt} = e^{V_{jt}} s_{0t}$. We want to find the values of coefficients in our utility function (Equation 1) that

minimize the distance between the estimated market shares s_{jt} and the observed market shares $\overline{S}_{jt}$. These values can be found simply by inverting $\overline{S}_{jt} = s_{jt}$, which can be done analytically. Thus we derive the following equation, which is used to test our hypotheses:

$$\ln \frac{\overline{S}_{jt}}{s_{0t}} = -\alpha \ln(\text{Price}_{jt}) + \sum_k \beta_k (\text{Search Characteristics}_{kjt}) + \sum_l \gamma_l (\text{Experience Characteristics}_{ljt}) - \sum_l \gamma_l' (\text{Experiene Characteristic}_{ljt}^2) + \sum_m \lambda_m (\text{Control Variables}_m) + \xi_{jt} \quad (2)$$

where s_{0t} denotes the estimated market share of outside goods (e.g., used cars and public transportation). Equation 2 brings us to an ordinary least-squares form, so we can address price endogeneity by running a two-stage instrumental variables regression.

3.2. Identification and Instruments

The dependent variable is aggregated at the year level, and some (but not all) of the independent variables in our model specification (Equation 2) could potentially be changed by firms reacting to demand shocks. We use an instrumental variables approach to address any endogeneity concerns arising from this potential simultaneity.

In the US automobile industry, most product features are determined well in advance of launch and cannot be changed within a given year—the time frame of our aggregation. We follow other papers that study this industry (e.g., Berry et al. 1995, Sudhir 2001, Balachander et al. 2009) in assuming that product features such as horsepower and size are exogenous.

That said, the product's price could still be biased because of simultaneity and/or measurement error. In response to demand shocks, auto manufacturers can change prices almost instantaneously by way of various promotions. Hence we use a set of well-known instruments to address price endogeneity (Berry et al. 1995): exogenous product characteristics, the sum of product characteristics of all other cars manufactured by the same firm, and the sum of product characteristics of all cars manufactured by other firms. Berry et al. showed that these instruments are optimal under Bertrand–Nash equilibrium. Moreover, when estimating our models, we test the validity of our instruments. In particular, we find that the Kleibergen–Paap *rk* Wald F-statistic is always greater than the critical value of the Stock–Yogo weak identification test and that the Hansen J-statistic is never significant; hence we do not reject the null of exogenous instruments.

We remark that fake or "sponsored" reviews are much less common in our context than in other industries owing to the dispersed incentive structure in the market for new US cars: car review platforms cater to a nationwide audience, but dealerships—which are the direct beneficiaries of any sale—have a geographically restricted target audience. So a dealership that unethically influences

customer reviews will have to bear all of the cost even though the odds are extremely low that any consumer affected thereby will purchase directly from the focal dealership. Moreover, car review platforms differ from other review platforms in neither identifying nor rewarding "power" reviewers; hence firms have less incentive to sponsor a reviewer, whose reach is necessarily unknown.

Finally, we examined reports provided by companies that specialize in detecting fake reviews—including Fakespot.com, Reviewmeta.com, and Trustwerty.com—in addition to academic research projects such as Reviewskeptic.com. All of these sites focus on reviews from Amazon, TripAdvisor, Yelp, and Apple's App Store; none of the platforms examined by these services review new car models. Given this additional evidence, it is reasonable to conclude that fake reviews are rare in our study's context.[2] We therefore assume that experience characteristics, such as ratings and sentiment, are exogenous.

3.3. Data and Variables

We build our database by merging two different sources of data. Publicly available customer review data were obtained from one of the leading online resources for automotive information. At the time of data collection, this automobile "infomediary" firm had one of the highest Web traffic rankings—for the automobile subcategory of consumer information—in the United States. Of all the visitors to this website, 78% are from the United States, and the demographic profile (in terms of gender, age, income, and education) of these visitors is similar to those visiting its two closest competitors (source: alexa.com). Car prices, specifications, and sales data were collected from WardsAuto, which is a leading provider of data on the automobile industry.

In our context, a "new car" is a noncommercial lightweight vehicle in its basic form (i.e., without any added options) that is newly released in the United States. Our customer review data identify 2,660 new cars (43 brands and 437 models) that have been reviewed. From the WardsAuto specification data, we observe 3,184 new cars (49 brands and 479 models) from 2002 to 2013. It is worth noting that the observations lost when these databases were merged—that is, owing to a lack of corresponding reviews—were not unpopular cars; most of the lost observations were of commercial trucks whose baseline model weight was less than 8,500 pounds and hence were categorized by WardsAuto as a "light" vehicle (but were not reviewed by consumers). Note also that exotic cars, such as the Rolls-Royce, were reviewed but not tracked by WardsAuto. Our final data set includes 2,392 observations of 416 car models manufactured by 40 different brands from 2002 to 2013 as well as 108,418 reviews that were aggregated at the model-year level.

[2] During 2017, there was one reported instance of a fake review in the automobile industry. However, it involved reviews of local car dealerships. This incident did not impact our data, which included only customer reviews of car models not dealerships (see https://www.getfivestars.com/blog/car-dealers-fined-3-6-million-fake-reviews-deceptive-practices/).

3.3.1. Dependent variable. As specified in Equation 2, the dependent variable in our estimation is the difference between (logged) market share and (logged) share of the outside good. We assume that a household requires at least one mode of transportation—either a new car, a used car, or public transportation—and that the market size is equal to the focal year's number of US households. Population demographic data were downloaded from the US Census Bureau's website. Market share is defined as the number of a model's new cars sold that year *divided by* the market size that year. The share of outside goods is determined by subtracting the sum of all new car sales that year from the market size and then dividing this difference by the market size.

3.3.2. Car characteristic variables. We use price, horsepower, size, and miles per dollar to characterize a car; in this we follow previous research that estimates demand in the automotive industry (Berry et al. 1995, Sudhir 2001, Petrin 2002, Balachander et al. 2009). We use the (log of the) *manufacturer's suggested retail price* (MSRP) of the base model, which correspond to the bottom-level trim package, because that price is the one most widely advertised across the country for a given car model. The MSRP differs from the actual *transaction price*, which depends not only on the trim level and the cost of additional options but also on negotiations with the car dealership. However, transaction prices are generally unknown (both to the econometrician and other consumers) because they are the outcomes of private negotiations. Consumers will enter negotiations and obtain transaction prices for only a selected few cars, since this process is time-consuming and requires substantial effort. It follows that, for most autos considered, consumers compare MSRPs rather than transaction prices. We adjust all car prices to 2014 US dollars via the inflation rate calculated using the Consumer Price Index, which was obtained from the US Bureau of Labor Statistics. The search characteristics referenced in Equation 2 are horsepower, size, and fuel efficiency. Our *HP_weight* variable is calculated by dividing a car's horsepower by its weight and then multiplying by 100. A vehicle's *Size* is calculated by multiplying its length by its height. We use miles per dollar (*MPD*) to capture fuel efficiency; this variable is calculated by dividing the car's city miles per gallon by the inflation-adjusted price for a gallon of gasoline. Historical gasoline price data were downloaded from the US Energy Information Administration's website. We also control for brand and year fixed effects.

3.3.3. Review characteristic variables. We use the review data to calculate experience characteristics, which consist of star ratings and the sentiment of text reviews. Our two key predictor variables are thus *Rating* and *Sentiment*. The variable *Rating*, which ranges from 1 to 5 stars, is a numerical average of star ratings received by a car model in the given year; *Sentiment* is a summary assessment of the comments written about a car model for a given year. More specifically, it is a variable that ranges from 0 to 1, which averages the sentiments conveyed in text reviews of a

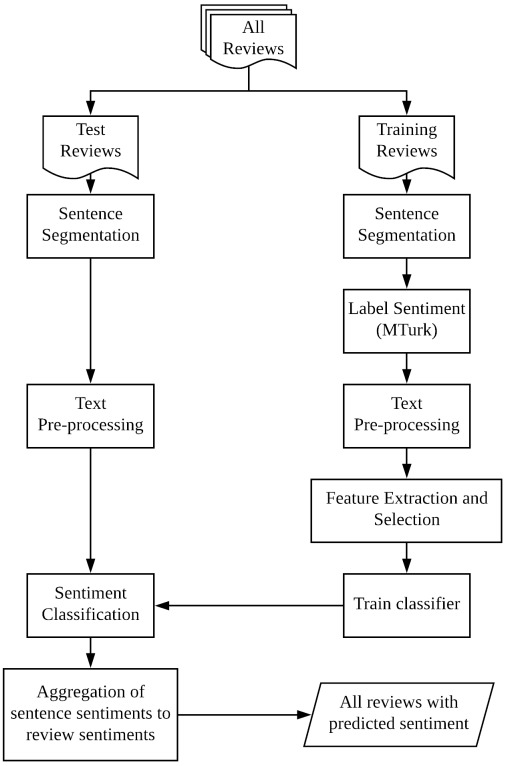
car model for a given year. Section 4 discusses in detail how we carried out the sentiment analysis of all product review comments to measure aggregated sentiment. In addition, we develop a set of control variables that capture various important features of customer reviews that could also affect consumer choices. We use only the reviews that were written in the year of a new model's release, since later reviews cannot influence that year's sales. Our *Num_review* variable is a count of the total number of reviews that a new car model received in the year of its launch. A given model has, on average, 45 reviews; the most popular model received more than 564 reviews. We control for the effect that rare negative signals may have on product demand. Maslowska et al. (2017) found that extremely positive ratings could actually depress demand, and they suggested that this dynamic may reflect credibility issues. Negative reviews are uncommon in the online context, so the distribution of star ratings tends to be left-skewed (Dellarocas 2003, Ghose and Ipeirotis 2006). We observe a similar trend in our data; only 3.6% of the ratings are fewer than 3 stars, and just 0.9% of the text reviews are classified as negative. Studies have shown that two-sided messaging—in which arguments both for and against an advocated position are presented—improves the credibility of the message's source (Hunt and Smith 1987, Eisend 2006). Hence we define *Has12* and *Has_negsent* as dummy variables that control for the diversity of opinions. The *Has12* indicator is set to 1 for car models that received at least one 1 or 2 star rating (and is set to 0 otherwise); *Has_negsent* is set to 1 for car models that received at least 1 text review with negative sentiments (and is set to 0 otherwise).

4. SENTIMENT ANALYSIS

Sentiment analysis is an automated, computational means of extracting the opinions contained in large volumes of unstructured text data, a task that would require a prohibitive amount of effort if undertaken manually by humans (Pang et al. 2002). We analyze the sentiment of the entire body of text—rather than examining just the most frequently occurring words—because analyzing the entire corpus allows us to develop a better understanding of the reviewer's overall attitude toward the product. This more comprehensive measure is also better suited for comparisons with ratings (which also capture a general view of the focal product) than would any compilation of feature-specific attitudes. The technique we employ for analyzing sentiment uses word choice, sentence structure, and context to determine a given comment's underlying connotation. Since there exist a variety of sentiment analysis methods and since applying supervised machine learning to analyze text is relatively new in the management literature and we are the first to apply it to sentiment analysis, we shall explain our method in some detail.

Sentiment classification methods can be broadly divided into two categories: those based on a lexicon and those based on machine learning (ML). Lexicon-based approaches rely on a pre-determined set of words or phrases and their associated connotations. The text is scoured for

Figure 1 Training the Classifier for Sentiment and Predicting Sentiment Orientation



occurrences of words in the lexicon, and the sentiment scores of matched words are averaged. The effectiveness of this approach depends strongly on the lexicon's coverage and appropriateness for the data's context, and it is generally lower than that of ML-based approaches. Moreover, lexicon-based techniques are not scalable for large data because they rely on a matching process. Given the substantial size of our sample text data—and because there is no car-specific lexicon—we adopt an ML-based method.

Machine learning enables the automatic selection and extraction of features that are relevant to the data domain via a set of "training data" of known sentiment. Other studies (Ghose et al. 2012, Netzer et al. 2012) have analyzed text by some other way of natural language processing (NLP). However, our method is closest to that of Lee et al. (2018), who implemented a supervised ML algorithm trained with data coded by humans—more specifically, using the Amazon Mechanical Turk (MTurk) crowdsourcing platform. Whereas Lee et al. were concerned with identifying the topics that are discussed, our goal is to classify the overall sentiment of a product review.

As shown in Figure 1, we split our review data into a training set and a test set. The training data set is a representative subset of the entire text review data ($\approx$1% of all the data). Both data sets undergo sentence segmentation and text pre-processing; at this stage, various NLP methods

are used to distill the text data into vectors of meaningful characteristics. For the training set only, we use the MTurk platform to obtain human input on classifying the sentiment of reviews at the sentence level before they are subject to text pre-processing. Then the sentiment classification algorithm, which we refer to as the *classifier*, learns from the training data how these vectors of characteristics should be grouped into the classes of positive and negative sentiment. Finally, the algorithm performs "sentiment prediction" on the rest of the data. Figure 1 illustrates the steps of this sentiment analysis, and in what follows we give more details of how the training data are collected and how the machine learning approach is applied.

4.1. Description of the ML-based approach

4.1.1. Creation of the training and test sets. We randomly sample all reviews to form a training data set. Our sampling procedure is such that the training set is representative with respect to the models, brands, and years of cars. Sampling at the review level is important to ensure that the training data will consist of entire reviews, which makes it more likely to be representative. The training set comprises 962 reviews and includes all 40 brands, all 12 years, and 287 models (out of 416). Our aim is to create a training set consisting of nearly 6,000 sentences; in the field of sentiment classification, typical training sets contain 3,000–5,000 sentences. The 962 reviews were broken down into 5,964 sentences via the segmentation method described next.

4.1.2. Sentence segmentation. We parse all reviews into sentences by first using the unsupervised method, proposed by Kiss and Strunk (2006), of detecting sentence boundaries. This unsupervised algorithm constructs a model for abbreviations, collocations, and words that start sentences; it then applies the model to find sentence boundaries. Next, we further deconstruct the sentences into clauses that contain similar sentiments by splitting each sentence at transition words (e.g., "but") that indicate contrast or exceptions. For our purposes, these sentence fragments are viewed as sentences.

4.1.3. Labeling the training set. We rely on MTurk, a crowdsourcing marketplace for simple tasks, to connect us with a large pool of participants who can assess the sentiment expressed in the training set's sentences. To ensure the quality of responses, we limit the participants to those who have previously completed more than 100 tasks on the platform and have an approval rating that exceeds 95%; we also include qualification questions and attention checks (see the Appendix for details). Furthermore, we limit participation to US residents (as verified by IP addresses) so that the sentiment coders are similar to consumers in our data. We ask the MTurk respondents to determine the sentiment of provided text on a 5-point Likert scale (extremely negative, somewhat negative, neutral, somewhat positive, extremely positive). Each respondent is asked 10 questions per task, and each sentence is evaluated by at least 10 respondents. We used the average of all

responses (where strongly negative $= -2$, somewhat negative $= -1$, neutral $= 0$, somewhat positive $= +1$, and strongly positive $= +2$) as our proxy of human-coded sentiment. Because sentiment is normally classified as simply being positive or negative, our classifier must be trained using binary inputs. Hence we label the sentiment of each sentence as positive if the average assessment is no less than 0 (and as negative otherwise). By this criterion, 25% of the sentences in the training set express a negative sentiment. We consider the sentiments so labeled as the true sentiment when training the classifier and (later) testing its performance.

4.1.4. Text pre-processing. All the sentences in both the training set and the test set are pre-processed before training the classifier and predicting sentiment class. Although the MTurk participants were provided with legible sentences, pre-processing is required of sentences if they are to be "read" efficiently by the ML algorithm. The purpose of this pre-processing step is to clear the sentences of "clutter", such as redundancies and frequently occurring words that convey no meaningful information. The major pre-processing steps are tokenization, removal of stop words, and stemming. Sentences are broken down into tokens, where a *token* is the smallest useful semantic unit. We then examine the tokens that occur most frequently and remove *stop words*—words (e.g., "the", "of ", "on") that do not affect the algorithm's accuracy but do increase complexity and thus reduce its speed. Finally, we use the Porter (1980) stemming algorithm to strip such suffixes as "-ing", "-es", and "-tion", thereby reducing each token to its stem.

4.1.5. Feature selection and extraction. By *features* we refer to useful pieces of information that are clear indicators of sentiment. We select the most relevant features by evaluating all tokens in the training set for their information gain—that is, for how much information a token's presence (or absence) contributes to classifying sentiment correctly. Feature selection is important because it not only reduces complexity but also eliminates redundant and "noisy" features that increase classification error and can lead to overfitting. As features we select the 500 tokens yielding the highest information gain. In the extraction process, each sentence in the *training* set is transformed into a 501×1 vector that indicates whether each of the 500 features is (or is not) present as well as the human-coded sentiment class; each sentence in the *test* set is transformed into a 500×1 vector indicating which of the 500 features is present.

4.1.6. Training the classifier. We consider a sentence, in the vector representation just described, to be a single data point. We train the algorithm by giving it data with the labeled sentiment from the training set. The algorithm then "learns" by building a model that best separates the data into their pre-determined sentiment classes. For this purpose we use a support vector machine (SVM), which looks for hyperplanes that linearly separate the data and then selects the hyperplane with the greatest distance to the nearest data points. This method reduces the risk of

overfitting the training set and also increases the test set's generalizability. According to Pang et al. (2002), the SVM method delivers results that are superior to those obtained by naïve Bayes or maximum entropy methods. We use an SVM with the radial basis function kernel, which allows the data to be projected to infinite dimensions and so increases the likelihood of linearly separating the data.

4.1.7. Sentiment prediction and aggregation. Finally, the trained algorithm is used to make predictions on sentences in the test set. Thus each new data point is assigned a positive or negative sentiment class membership based on its position relative to the optimally separating hyperplane determined via the training set. When predicting the sentiment, we use Platt scaling to transform predicted binary classifications into estimated class membership probabilities. This scaling technique yields probability estimates by way of the logistic transformation $P(y = 1|x = 1) = 1/(1 + exp(Af(x) + B))$; here y is a binary class label of a classifier, x is the input for that classifier, and A and B are scalar parameters that are learned and optimized with the aid of training data. Predictions are made at the sentence level, and we average the predicted sentiment values for all of a given review's sentences. The sentiments of all reviews for the focal car are then averaged so that each review is given equal weight.

4.2. Performance of the ML-based approach

We evaluate the sentiment classifier's performance by comparing the predicted binary sentiment classification against actual sentiment, which in our study is the sentiment determined by MTurk responses. Since we have actual sentiment only for the training set, we perform fivefold cross-validation on that data set as follows. After dividing the training set into five "folds", we: (i) hold out one fold and use the remaining four folds to train the algorithm; and (ii) use the trained classifier to make predictions on the held-out fold. We repeat these two steps for each fold and thus derive predictions for the entire training set. These predictions are representative of predictions that the algorithm would make (after it was trained) on the entire set—that is, instead of on four fifths of the training set. We use both the predicted and actual sentiment classifications to calculate four performance metrics: accuracy, precision, recall, and F1; each of these metrics is described next.

Accuracy, precision, recall, and F1 are various ratios involving true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). A true positive occurs when the classifier correctly identifies a positive statement, and a false positive occurs when the classifier incorrectly identifies a negative statement as positive; corresponding statements hold for true negatives and false negatives. *Accuracy* is the ratio of correct predictions to total observations, (TP + TN) / (TP + TN + FP + FN), and it serves as a simple measure of performance. *Precision* is the ratio

Table 1 Evaluating the performance of sentiment classifiers

		Recall		Precision		F1	
Method	**Accuracy**	**0**	**1**	**0**	**1**	**0**	**1**
Machine learning–based	0.8040	0.2706	0.9715	0.7585	0.8088	0.3965	0.8825
Lexicon-based	0.7360	0.6384	0.7668	0.4632	0.8706	0.5368	0.8154

of correct positive (resp. negative) predictions to total positive (resp. negative) observations: TP / (TP + FP) or TN / (TN + FN), respectively. *Recall* is the ratio of correct positive or negative predictions to all observations in its actual class: TP / (TP + FN) or TN / (TN + FP). The F1 score is a weighted average of precision and recall:

$$F1 = \frac{2(precision \times recall)}{precision + recall}$$

We benchmark the performance of the ML-based approach against the standard lexicon-based approach, specifically the widely used sentiment lexicon of Hu and Liu (2004). Recall that, with a lexicon-based approach, the words in each sentence are matched with those in the lexicon; then a word that matches a positive (resp. negative) lexicon word is assigned a score of +1 (resp. −1). The sentiment score of a sentence is the sum of the scores of matched words, and a sentence is classified as having positive (resp. negative) sentiment if that score is no less than 0 (resp. otherwise). Sentences that contain no matched sentiment words are randomly classified as positive or negative.

Table 1 summarizes the performance of all classifiers. The reported values establish that the ML-based method's accuracy, (i.e., the overall performance measure) exceeds that of the lexicon-based method by 7.8%. Similarly to previous research that has examined online feedback and reviews (Dellarocas 2003, Godes and Mayzlin 2004, Chevalier and Mayzlin 2006), our review data is overwhelmingly positive: 89.5% of the reviews have star ratings greater than 4 (out of 5), and 75% of the sentences in our training set were labeled as positive. The low recall we observe with our ML-based approach is an issue known to be associated with skewed data. Yet with regard to the recall and precision of negative sentiments, we remark that the ML-based approach is more conservative in labeling a statement as negative; however, when it does label a statement as negative, its accuracy is better than when the lexicon-based method is used.

Note that, although the predicted *binary* classifications were used to compare the performance of these two approaches, the estimated *probabilities of* class membership were used to calculate the sentiment variable for the various car models. The class membership probabilities, which were estimated using Platt scaling (see Section 4.1.7), allow us to use information on the classifier's certainty in making the prediction. A probability close to 1 or 0 indicates that the classifier was

Table 2 Summary statistics of variables (N=2,392 car model-year observations)

			Correlation Matrix								
Variable	**Mean**	**Std. Dev.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1. DV	-8.166	1.388	1								
2. Price	10.093	0.470	−0.510*	1							
3. HP_weight	5.928	1.960	−0.333*	0.381*	1						
4. Size	11.833	2.047	0.173*	0.173*	−0.247*	1					
5. MPD	9.745	5.069	0.126*	−0.453*	0.094*	−0.326*	1				
6. Num_review	45.325	59.935	0.428*	−0.303*	−0.096*	−0.075*	0.326*	1			
7. Has12	0.558	0.497	0.422*	−0.307*	−0.170*	0.024	0.194*	0.407*	1		
8. Rating	4.521	0.322	−0.123*	0.083*	0.150*	−0.169*	0.125*	0.167*	−0.353*	1	
9. Has_negsent	0.623	0.485	0.400*	−0.273*	−0.121*	−0.033	0.235*	0.412*	0.502*	−0.136*	1
10. Sentiment	0.765	0.063	−0.072*	−0.019	0.059*	−0.079*	0.196*	0.150*	−0.152*	0.595*	−0.255*

* Correlation coefficient is significant at p=0.05 level.

extremely certain in labeling the sentence as positive or negative, whereas a probability close to 0.5 indicates that the classifier was fairly uncertain about its prediction. Thus using these probabilities—instead of the binary predictions—enables us to exploit more of the classifier-generated information.

5. RESULTS

Summary statistics and pairwise correlations of our variables are shown in Table 2. Table 3 summarizes the regression results. Model [1] reports baseline results with product characteristics as controls, and Model [2] includes both product and review characteristics as controls. We note that the estimated coefficients for both diversity controls (*Has12* and *Has_negsent*) are positive and significant, which is consistent with the findings of prior studies (Hunt and Smith 1987, Eisend 2006). So even though the rare extremely negative reviews may lower (incrementally) the average rating or sentiment, they seem to play an important role in establishing credibility and ultimately have a positive effect on demand. That said, all our results are robust to the exclusion of these two controls for review characteristics.

Model [3] in Table 3 reveals that the average star rating has a positive and significant effect on demand (0.393, p ¡ .01). This result confirms findings from extant research (Godes and Mayzlin 2004, Chevalier and Mayzlin 2006, Dellarocas et al. 2007, Forman et al. 2008, Ghose and Ipeirotis 2011, Lehman et al. 2014, Luca 2016). Model [4] shows that the average sentiment has a positive and significant effect (1.114, p ¡ .05). In Models [5] and [6], we assume a quadratic relationship between demand and respectively ratings and sentiment. Model [7] includes the effects of both ratings and sentiment. We find that the coefficient for linear ratings and linear sentiment remains consistently

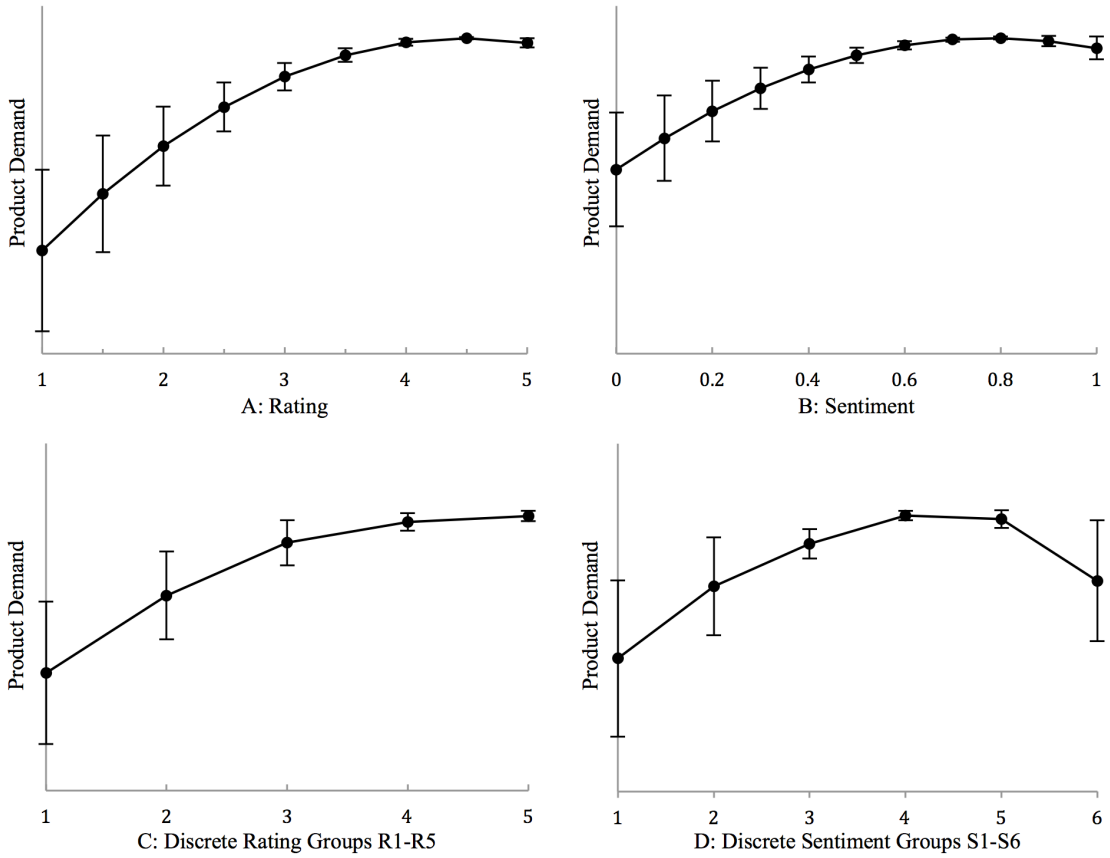
Table 3 Regression models predicting product demand (N = 2,392 car model-year observations)

Variable	[1] Car controls	[2] Review controls	[3] Linear ratings	[4] Linear sentiment	[5] Quadratic ratings	[6] Quadratic sentiment	[7] Main effects	[8] Semi-parametric rating	[9] Semi-parametric sentiment	[10] Interaction effect
Price	–1.368** (0.544)	–1.480*** (0.476)	–1.379*** (0.464)	–1.370*** (0.463)	–1.361*** (0.462)	–1.364*** (0.468)	–1.321*** (0.460)	–1.309*** (0.470)	–1.267*** (0.461)	–1.349*** (0.457)
HP_weight	–0.0958** (0.0449)	–0.0410 (0.0368)	–0.0516 (0.0360)	–0.0518 (0.0359)	–0.0494 (0.0355)	–0.0493 (0.0363)	–0.0520 (0.0354)	–0.0543 (0.0364)	–0.0547 (0.0355)	–0.0496 (0.0353)
Size	0.0968*** (0.0261)	0.113*** (0.0215)	0.116*** (0.0217)	0.117*** (0.0217)	0.116*** (0.0215)	0.115*** (0.0214)	0.116*** (0.0214)	0.117*** (0.0217)	0.118*** (0.0214)	0.117*** (0.0213)
MPD	–0.000943 (0.0301)	–0.0243 (0.0259)	–0.0192 (0.0253)	–0.0180 (0.0252)	–0.0180 (0.0253)	–0.0179 (0.0254)	–0.0158 (0.0252)	–0.0152 (0.0256)	–0.0129 (0.0253)	–0.0174 (0.0250)
Num_review		0.00710*** (0.000484)	0.00655*** (0.000492)	0.00654*** (0.000491)	0.00684*** (0.000507)	0.00692*** (0.000480)	0.00684*** (0.000505)	0.00671*** (0.000495)	0.00684*** (0.000506)	0.00681*** (0.000502)
Has12		0.279*** (0.0457)	0.399*** (0.0470)	0.382*** (0.0478)	0.338*** (0.0467)	0.282*** (0.0446)	0.315*** (0.0473)	0.345*** (0.0459)	0.309*** (0.0467)	0.312*** (0.0469)
Has_negsent		0.449*** (0.0464)	0.476*** (0.0456)	0.524*** (0.0494)	0.452*** (0.0452)	0.506*** (0.0477)	0.479*** (0.0484)	0.469*** (0.0484)	0.493*** (0.0460)	0.461*** (0.0480)
Rating			0.393*** (0.0858)	0.276** (0.109)	3.213*** (0.749)		2.420*** (0.798)	$R_1,\ldots,R_5$ [a]	1.928*** (0.740)	5.420*** (1.138)
Rating^2					–0.353*** (0.0917)		–0.268*** (0.0979)	$R_1,\ldots,R_5$ [a]	–0.216** (0.0912)	–0.605*** (0.134)
Sentiment				1.114** (0.565)		11.67** (4.744)	8.281** (4.119)	8.662** (4.395)	$S_1,\ldots,S_6$ [b]	11.57*** (2.781)
Sentiment^2						–7.237** (3.269)	–5.289* (2.854)	–5.732* (3.015)	$S_1,\ldots,S_6$ [b]	–7.372*** (1.969)
Rating × Sentiment										3.069*** (0.742)
Constant	4.985 (5.550)	5.029 (4.920)	2.078 (4.687)	1.586 (4.685)	–3.586 (5.355)	–0.973 (5.273)	–5.423 (5.363)	–1.463 (5.151)	–2.854 (5.328)	–13.09** (5.708)
R-squared	0.516	0.649	0.654	0.655	0.659	0.658	0.662	0.663	0.664	0.666

Notes: Reported values are correlation coefficients with robust standard errors in parentheses. All regressions incorporate brand and year fixed effects.
[a] R_1 through R_5 are indicator variables: R_j ($j = 2,\ldots,5$) is set to 1 for ratings within the range ($j + 1$, $j + 1.5$] and is set to 0 otherwise; $R_1 = 1$ if the rating does not exceed 3 (out of a possible 5) stars and is 0 otherwise. We group low ratings together because they account for only 0.5% of the data ($N = 12$). Coefficients (with standard errors in parentheses): $R_2 = 0.669$* (0.377), $R_3 = 1.124$*** (0.312), $R_4 = 1.305$*** (0.316), $R_5 = 1.358$*** (0.317).
[b] S_1 through S_6 are indicator variables: S_j ($j = 2,\ldots,6$) is set to 1 for sentiment within the range ($j/10 + 0.3$, $j/10 + 0.4$] and is set to 0 otherwise; $S_1 = 1$ if the sentiment does not exceed 0.5 (on a scale of 0 to 1) and is 0 otherwise. We group low sentiments together because they account for only 0.5% of the data ($N = 11$). Coefficients (with standard errors in parentheses): $S_2 = 0.623$* (0.376), $S_3 = 0.990$*** (0.345), $S_4 = 1.233$*** (0.346), $S_5 = 1.200$*** (0.348), $S_6 = 0.670$ (0.429).

*$p < .10$, **$p < .05$, ***$p < .01$

Figure 2 Marginal Effects of Continuous and Discrete Rating and Sentiment on Product Demand

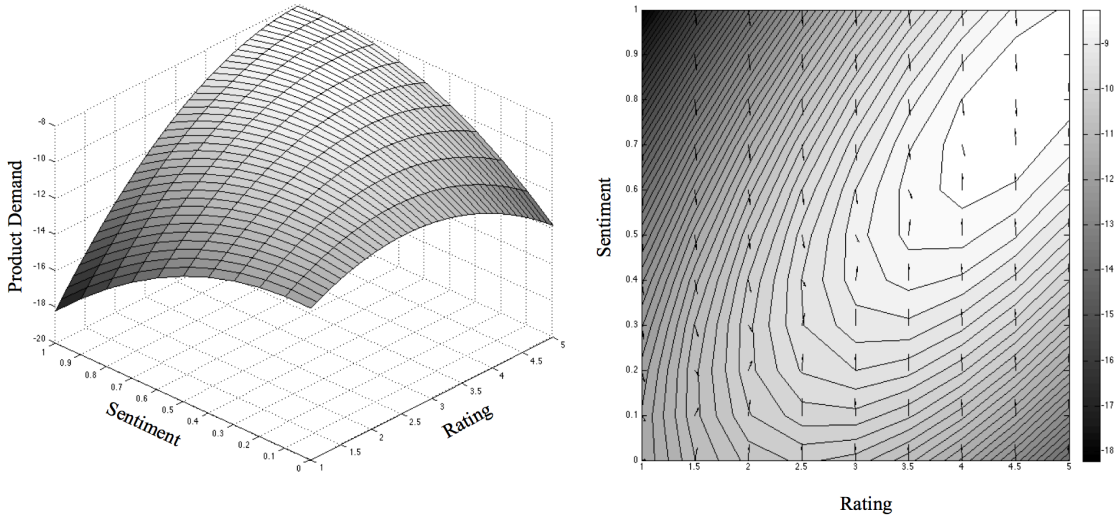


positive and significant whereas the quadratic terms are *negative* and significant, which indicates a decreasing sensitivity to increases in either ratings or sentiment (in line with H1a and H1b). Panel A in Figure 2 plots the marginal effects of a product's star rating on its demand, as specified in Model [7]; Panel B similarly plots the marginal effects of text sentiment on demand (also based on Model [7]). These graphs illustrate how the effect of a small increase in ratings/sentiment is much greater for cars with low than with high ratings or sentiment.

In order to avoid imposing any functional form on the relationship between demand and ratings or sentiment, we estimated semi-parametric models (Models [8] and [9] in Table 3). The marginal effect of ratings is illustrated in Figure 2's Panel C; this graph clearly shows decreasing sensitivity to an increased rating and is consistent with results from Model [7], which strongly supports H1a. Panel D similarly shows decreasing sensitivity to increases in sentiment for all but extremely high values of sentiment. Thus we find tentative support for H1b and so, to investigate further the effect of such high values, we examine the model with interaction effects

We test Hypothesis 2 in Model [10], which includes a term for the interaction between rating and sentiment. That interaction term was mean-centered because the *Ratings* variable is never equal to 0 (it ranges from 1 to 5). The coefficients for both the linear and quadratic rating and

Figure 3 Joint Effect of Review Rating and Sentiment on Product Demand



sentiment terms are consistent with results from Models [5]–[9]: positive for the linear term and negative for the quadratic term, which indicates a decreasing sensitivity to increases. The coefficient for the interaction term (*Ratings* × *Sentiment*) is positive and significant, so the marginal effect of ratings is increasing in sentiment. This finding supports H2; ratings and sentiment have a complementary effect on product demand, which means that sentiment can compensate for the tendency of consumers to discount highly positive ratings.

To develop a better understanding of this intricate complementary effect, in Figure 3 we plot the contours of the three-dimensional (sentiment × rating × demand) space and overlay the normalized gradient vectors. Along the diagonal that runs from the graph's bottom left to its top right are the cases for which ratings and sentiment are in agreement; here we show that the effect of ratings is monotonically increasing at a decreasing rate, which provides further support for H1a. The same diagonal also reveals that the effect of sentiment is monotonically increasing at a decreasing rate, which offers additional support for H1b. It is noteworthy that positive sentiments compensate for the tendency to discount highly positive star ratings.

Above and below this diagonal are instances of disagreement between ratings and sentiment. In cases where these two signals of quality disagree, the plot shows that the demand decreases. The reason is that, when there is disagreement (i.e., when one signal is high and the other is low), the discounting effect driven by the high signal is strong yet the complementarity effect is nonexistent for the low–high pairing. In other words, the tendency of potential customers to discount highly positive star ratings is amplified in the presence of negative sentiments. Note also that the declining effect of sentiment on demand at extremely high values of sentiment noted earlier (from Panel D in Figure 2) can be explained by the disagreement in signals, which is consistent with our argument

that reducing a signal's credibility can dampen its effect (Mudambi and Schuff 2010, Maslowska et al. 2017).

Robustness tests

Table 4 is devoted to alternative model specifications that test the robustness of our results. Overall, those results are robust to alternative model and variable specifications. Model [11] reports results when we do not control for price endogeneity. Both the sign and the statistical significance of each covariate's coefficient are consistent with Model [10], which does account for price endogeneity.

In Model [12] we check for the possibility that long-run average ratings are biased because reviewers exhibit herding behavior. Except for a product's very first reviewer, all reviewers see the (prominently displayed) average star rating *before* writing their own review. This sequence dictates that we consider the possibility of herding behavior. Several experiments have shown that participants tend to revise their quantitative estimates after seeing the estimates given by other individuals. Thus a participant's final assessment tends to be a weighted average: 70% is due to their own initial estimate and 30% to the estimates of others (Harvey and Fischer 1997, Yaniv 2004, Soll and Larrick 2009). Hence in this robustness test we suppose that herding has occurred and adjust our Ratings variable by assuming that all ratings, except for the first, reflect not only the reviewer's "true" ex post quality estimate (70% weight) but also the average of all prior reviews (30% weight); we then average only the "true" quality estimates. Model [12] reports the results when we use this adjusted average rating, and it shows that our findings are robust to possible herding effects. Although we do not include in the paper, we similarly calculate new averages while assuming various levels of herding behavior—namely, weights of 10% to 90%, in 10% increments, given to prior reviews—and obtain results consistent with the no-herding condition as long as the weighting does not exceed 60%.

Model [13] in Table 4 tests the robustness of our results to the presence of car models with "too few" product reviews. Li and Hitt (2008) showed that early reviews are systematically positively biased but then "settle" over time. To ensure that the long-run average is indeed representative of true ex post quality, we omit car models for which there are not enough reviews; the average opinion of cars with only a few reviews may depend too strongly on potentially biased early reviews and so may have not yet settled. We define *settled* reviews as a set whose cumulative average reaches and remains within 5% of the long-run average. In our data, cars that have settled according to this criterion are reviewed 6.19 times on average. In this robustness test, we therefore omit from the analysis all cars with fewer than seven reviews and find that the results are consistent with our main findings.

Next we investigate how sentiment measurement error might affect our results. More specifically, we compare the predicted sentiment values from the *supervised* ML algorithm (Model [10] in Table

Table 4 Results of selected robustness tests

Variable	[11] Exogenous price	[12] Herding at 30%	[13] Settled reviews only	[14] Unsupervised ML sentiment	[15] Subjectivity control	[16] SD as diversity control	[17] 5 most recent sentiment	[18] 5 most negative sentiment
Price	–1.460***	–1.351***	–1.188**	–1.345***	–1.359***	–1.534***	–1.360***	–1.174**
	(0.0872)	(0.460)	(0.481)	(0.464)	(0.458)	(0.473)	(0.458)	(0.473)
HP_weight	–0.0415***	–0.0496	–0.0567*	–0.0505	–0.0489	–0.0400	–0.0486	–0.0608*
	(0.0134)	(0.0355)	(0.0344)	(0.0357)	(0.0354)	(0.0353)	(0.0353)	(0.0363)
Size	0.121***	0.117***	0.116***	0.121***	0.118***	0.133***	0.117***	0.110***
	(0.0124)	(0.0214)	(0.0252)	(0.0214)	(0.0213)	(0.0219)	(0.0213)	(0.0217)
MPD	–0.0232**	–0.0178	–0.0140	–0.0178	–0.0179	–0.0232	–0.0183	–0.00974
	(0.00931)	(0.0252)	(0.0243)	(0.0253)	(0.0251)	(0.0258)	(0.0251)	(0.0258)
Num_review	0.00677***	0.00680***	0.00684***	0.00672***	0.00681***	0.00827***	0.00680***	0.00672***
	(0.000461)	(0.000496)	(0.000530)	(0.000499)	(0.000502)	(0.000563)	(0.000501)	(0.000522)
Has12	0.310***	0.311***	0.156***	0.317***	0.312***		0.310***	0.279***
	(0.0471)	(0.0457)	(0.0482)	(0.0467)	(0.0468)		(0.0465)	(0.0465)
Has_negsent	0.458***	0.460***	0.352***	0.434***	0.462***		0.444***	0.273***
	(0.0477)	(0.0481)	(0.0529)	(0.0447)	(0.0480)		(0.0464)	(0.0523)
Rating	5.372***	4.800***	15.53***	5.879***	5.423***	11.18***	4.777***	4.101***
	(1.132)	(0.899)	(3.439)	(0.863)	(1.127)	(1.575)	(1.093)	(0.815)
$Rating^2$	–0.598***	–0.538***	–1.768***	–0.651***	–0.605***	–1.268***	–0.528***	–0.447***
	(0.132)	(0.107)	(0.393)	(0.103)	(0.133)	(0.186)	(0.130)	(0.0991)
Sentiment	11.56***	11.51***	48.30***	50.72***	11.32***	47.79***	12.43***	6.544***
	(2.813)	(2.869)	(14.79)	(11.83)	(2.837)	(7.892)	(2.992)	(1.534)
$Sentiment^2$	–7.368***	–7.330***	–30.80***	–41.38***	–7.181***	–31.64***	–8.315***	–5.997***
	(1.992)	(2.028)	(9.644)	(9.788)	(2.021)	(5.212)	(2.103)	(1.234)
Rating × Sentiment	3.065***	2.696***	11.60***	9.146***	3.053***	7.984***	2.520***	1.767***
	(0.759)	(0.633)	(3.153)	(1.447)	(0.734)	(1.425)	(0.681)	(0.531)
Subjectivity					–0.140			
					(0.408)			
Rating_SD						0.200		
						(0.132)		
Sentiment_SD						1.293**		
						(0.576)		
Constant	–11.86***	–11.58**	–50.67***	–25.24***	–12.86**	–37.00***	–11.71**	–9.066
	(2.675)	(5.356)	(13.23)	(7.167)	(5.783)	(7.505)	(5.491)	(5.649)
Observations	2,392	2,392	1,900	2,392	2,392	2,276	2,392	2,392
R-squared	0.666	0.665	0.666	0.667	0.666	0.657	0.666	0.666

Notes: Reported values are correlation coefficients with robust standard errors in parentheses. All regressions incorporate brand and year fixed effects.

SD = standard deviation.

*$p < .10$, **$p < .05$, ***$p < .01$

3) to those from the *unsupervised* ML algorithm (Model [14] in Table 4). Although accuracy of the unsupervised ML algorithm's performance (73.6%) is less than that of the supervised ML algorithm (80.4%), each covariate's coefficient has the same sign and statistical significance as in our main findings.

In Models [15] and [16], we examine other factors that could affect the credibility of text reviews. In Model [15] we create a *Subjectivity* variable (based on the unsupervised ML approach) to distinguish subjective statements from objective statements. We expect that highly subjective reviews

will be less credible and that reviews containing many subjective statements will have a negative effect on demand. The values reported for this model indicate that the coefficient for *Subjectivity* is negative but not statistically significant. The effect of both linear and quadratic sentiment is consistent with Model [10] (in Table 3), which suggests that credibility issues arise more from extremely positive reviews than from extremely subjective ones. Model [16] replaces the diversity controls *Has12* and *Has_negsent* with standard deviations (SDs) of ratings and sentiments, where a high SD is indicative of opinions that are relatively more diverse. The correlation coefficients are positive for the standard deviations of both ratings and sentiment (as were those for *Has12* and *Has_negsent*), but the effect of the ratings SD is not significant.

For popular cars with hundreds of reviews, it is unreasonable to expect that a potential customer will read all of them. We mimic two popular sorting behaviors in this context: sorting by the most recent reviews and sorting by lowest star ratings. The default sorting option for our data is by most recent, under which five reviews are shown per web page[3]. Hence for this test we assume that, at any given time, the five most recent reviews are more likely to be read than the remaining reviews; we then create a variable for the yearly average of the moving average of five most recent reviews' sentiments. In Model [17], we substitute this variable for the baseline version (i.e., for the average sentiment of *all* reviews) and derive results that are consistent with those in Model [10]. We also consider the possibility that some people deliberately seek out the most negative reviews in order to identify potential deal-breakers. So in Model [18], we use the yearly average of the moving average of five most negative reviews' sentiments and then confirm that our results are robust to this alternative.

6. DISCUSSION

Understanding how consumers perceive product quality—and how this perception affects product demand—are of great interest to all firms and especially to those providing complex products or services, whose quality is difficult to assess. Given the wealth of product reviews on e-commerce websites, review platforms, and microblogs, operations management scholars and practitioners have realized that product design and pricing strategies must be adjusted to account for the social learning about product quality that is enabled by such online product review platforms (Mckinsey 2010, Yu et al. 2016, Crapis et al. 2017, Feldman et al. 2018). Despite these efforts to understand the role played by the increasing presence of online product reviews in various aspects of the firm's strategy, we have only a limited understanding of the precise mechanisms by which product reviews influence product demand. Addressing this gap is important because such understanding will enable

[3] Since the time our data were collected, the review platform has undergone a major design change. The default sorting option now is by "most helpful".

the firm to gain knowledge not only of how its existing customers have used its products and assessed their quality ex post but also of how its potential future customers perceive its products' quality as compared with other products in the market. These undertakings require the firm to recognize that product reviews are expressed in both quantitative and qualitative forms and that both types of assessment are related—in complementary and nontrivial ways—to expected product demand.

Our study is the first to test explicitly whether (or not) star ratings and text reviews convey the same information and to study the interplay between these two reviewing modes and how they affect product demand. We demonstrate empirically that the effect of extremely positive ratings and text reviews is diminished by skepticism yet also that, when ratings and review sentiment are in accord, such credibility problems are mitigated. An underspecified model that fails to consider the interaction between star ratings and text sentiment might incorrectly determine that the demand for highly rated products is low. In fact, our results indicate that the demand for highly rated products is low only when those extremely positive ratings are *not* supported by equally positive text reviews, and our analysis of such interactions underscores the importance of considering review sentiment and star ratings simultaneously.

Drawing on the dual-process theory of cognitive psychology, we develop a theoretical argument for why star ratings and text reviews may capture different types of information: star ratings reflect the output of System 1 thinking whereas text review sentiments reflect the output of System 2 thinking. From the perspective of potential customers, star ratings (resp., review sentiments) analogously trigger System 1 (resp., System 2) thinking. This has allowed us to conceptually ground the notion of ex post quality onto fundamentals of individual decision-making.

An interesting practical implication arises from our drawing on the dual-process theory. In light of this literature, it seems likely that text reviews provide a more complete signal of quality than do star ratings. In other words: since System 2 (which depends on the reviewer's cognitive ability and resources) can correct for biases due to System 1, it follows that text sentiment is more likely to capture how the analytical aspect of an individual views the focal product's quality. Yet because of how these text reviews are displayed, it requires more effort to digest this "better" information—it is far more difficult to wade through pages of a text review than to see how many stars the product has garnered. Hence product review platforms should ensure that text reviews are presented in a way that is easier to consume. For example, some review platforms now present text reviews that incorporate a curated list of "pros" and "cons" derived from the full text reviews. Such alternatives offer the opportunity for future research on how different review platform designs affect both product perception and product demand.

The perspective of dual-process theory yields another implication of interest related to sentiment analysis. Some sentiment analysis applications have incorrectly assumed that ratings and sentiment are simply two forms of the same assessment and used ratings to train their classifier. However, our findings strongly suggest that sentiments and ratings result from two distinct cognitive processes. These results corroborate our approach to sentiment analysis, whereby sentiments are labeled using multiple human inputs independently of star ratings. We conclude that the underlying sentiment of product reviews can be truly understood only if star ratings are *not* used to train the review text classifier.

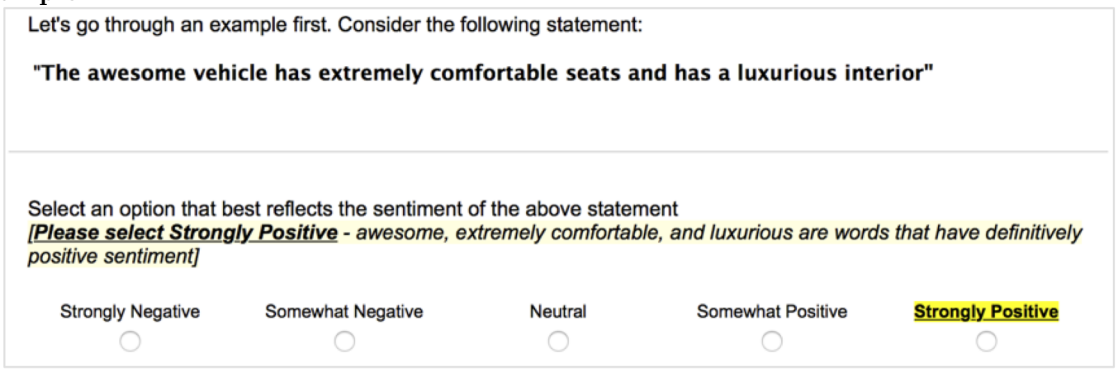
Despite the robustness of our findings, there are some limitations to our work. Advances in natural language processing and ML-based text analysis have enabled this study. Although we have shown that our findings are robust to small differences in algorithm performance—80.4% accuracy for supervised learning versus 73.5% accuracy for unsupervised learning—it is still possible that a more significant improvement in performance may affect our findings. Whereas today's state-of-the-art *topic* classification algorithms have an accuracy of about 95%, their *sentiment* classification counterparts exhibit only 80% accuracy; hence the latter could see significant improvement in the near future.

Another factor worth considering is that our data do not rule out the possibility of reviewers changing their star ratings after the text review is written. The dual-process literature implies that initial perceptions are not easily changed because System 1 thinking is essentially a "black box" process: individuals are seldom aware of exactly how they arrive at a conclusion. Even so, this possibility is an intriguing avenue for future work; experiments could examine how changing the order (i.e., comments *before* ratings) in which these review components are solicited (and displayed) affects product perception and demand. Writing the comments first forces the reviewer to thereby read her System 2 thinking, which may allow star ratings to capture an updated response that corrects for the biases from System 1.

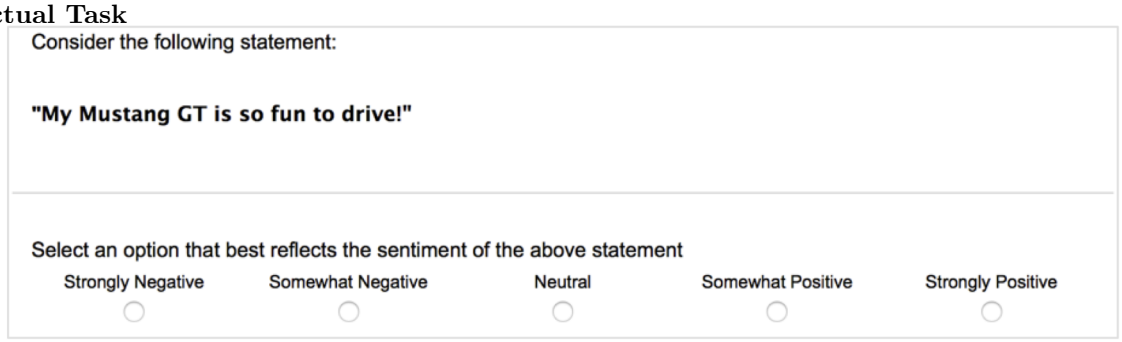
In positing that product reviews capture ex post quality, we have argued that product reviews are an excellent resource for understanding how consumers perceive—both intuitively and rationally—a product's quality. For manufacturers and service providers alike, our findings call for closely monitoring both the quantitative and qualitative aspects of product reviews to understand the extent to which manufacturer intended quality is actually perceived by consumers.

Appendix. MTurk details

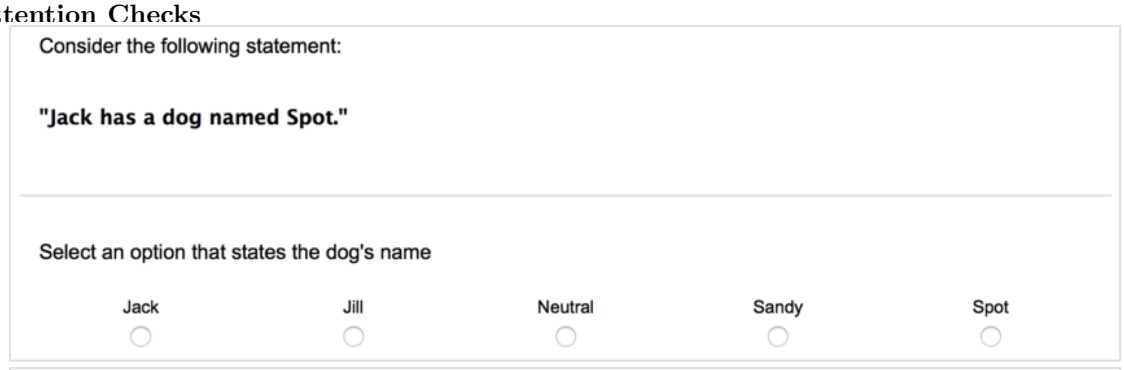
Example

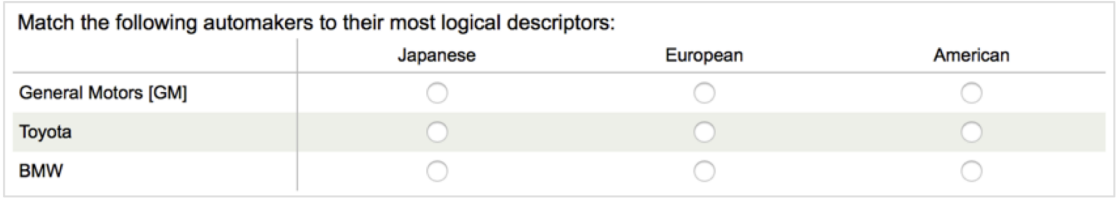


Actual Task



Attention Checks





MTurk Responses

Status	Count	%
Bounced*	852	1.4
Failed attention checks	8,490	13.8
Completed	52,030	84.8
Total	**61,372**	**100**

* "Bounced" MTurkers are those who exited the survey early (i.e., without obtaining a completion code).

Acknowledgments

References

Archak N, Ghose A, Ipeirotis PG (2011) Deriving the pricing power of product features by mining consumer reviews. *Management science* 57(8):1485–1509.

Armstrong JS (2001) Combining forecasts. *Principles of forecasting*, 417–439 (Springer).

Avery C, Resnick P, Zeckhauser R (1999) The market for evaluations. *American Economic Review* 89(3):564–584.

Balachander S, Liu Y, Stock A (2009) An empirical analysis of scarcity strategies in the automobile industry. *Management Science* 55(10):1623–1637.

Berry S, Levinsohn J, Pakes A (1995) Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society* 841–890.

Bickart B, Schindler RM (2001) Internet forums as influential sources of consumer information. *Journal of interactive marketing* 15(3):31–40.

Chetty R, Looney A, Kroft K (2009) Salience and taxation: Theory and evidence. *American economic review* 99(4):1145–77.

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

Clemen RT (1989) Combining forecasts: A review and annotated bibliography. *International journal of forecasting* 5(4):559–583.

Clemen RT, Winkler RL (1986) Combining economic forecasts. *Journal of Business & Economic Statistics* 4(1):39–46.

Crapis D, Ifrach B, Maglaras C, Scarsini M (2017) Monopoly pricing in the presence of social learning. *Management Science* 63(11):3586–3608, ISSN 1526-5501, URL `http://dx.doi.org/10.1287/mnsc.2016.2526`.

Dellarocas C (2003) The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science* 49(10):1407–1424.

Dellarocas C (2006) Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management science* 52(10):1577–1593.

Dellarocas C, Zhang XM, Awad NF (2007) Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive marketing* 21(4):23–45.

Deutsch M, Gerard HB (1955) A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology* 51(3):629.

Duan W, Gu B, Whinston AB (2008) Do online reviews matter?—an empirical investigation of panel data. *Decision support systems* 45(4):1007–1016.

Economides N, Seim K, Viard VB (2008) Quantifying the benefits of entry into local phone service. *the RAND Journal of Economics* 39(3):699–730.

Eisend M (2006) Two-sided advertising: A meta-analysis. *International Journal of Research in Marketing* 23(2):187–198.

Epstein S (1994) Integration of the cognitive and the psychodynamic unconscious. *American psychologist* 49(8):709.

Evans JSB, Stanovich KE (2013) Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science* 8(3):223–241.

Feldman P, Papanastasiou Y, Segev E (2018) Social learning and the design of new experience goods. *Management Science* .

Festinger L (1954) A theory of social comparison processes. *Human relations* 7(2):117–140.

Forman C, Ghose A, Wiesenfeld B (2008) Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research* 19(3):291–313.

Ghose A, Ipeirotis PG (2006) Designing ranking systems for consumer reviews: The impact of review subjectivity on product sales and review quality. *Proceedings of the 16th annual workshop on information technology and systems*, 303–310.

Ghose A, Ipeirotis PG (2011) Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering* 23(10):1498–1512.

Ghose A, Ipeirotis PG, Li B (2012) Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science* 31(3):493–520.

Gilbert DT (1991) How mental systems believe. *American psychologist* 46(2):107.

Godes D, Mayzlin D (2004) Using online conversations to study word-of-mouth communication. *Marketing science* 23(4):545–560.

Graham JR (1996) Is a group of economists better than one? than none? *Journal of Business* 193–232.

Harvey N, Fischer I (1997) Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes* 70(2):117–133.

Hu M, Liu B (2004) Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177 (ACM).

Hunt JM, Smith MF (1987) The persuasive impact of two-sided selling appeals for an unknown brand name. *Journal of the Academy of Marketing Science* 15(1):11–18.

Kahneman D (2011) *Thinking, fast and slow*, volume 1 (Farrar, Straus and Giroux New York).

Kahneman D, Frederick S (2002) Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment* 49:81.

Kim Y, Telang R, Vogt WB, Krishnan R (2010) An empirical analysis of mobile voice service and sms: a structural model. *Management Science* 56(2):234–252.

Kiss T, Strunk J (2006) Unsupervised multilingual sentence boundary detection. *Computational Linguistics* 32(4):485–525.

Kök AG, Fisher ML, Vaidyanathan R (2015) Assortment planning: Review of literature and industry practice. *Retail supply chain management*, 175–236 (Springer).

Larrick RP, Soll JB (2006) Intuitions about combining opinions: Misappreciation of the averaging principle. *Management science* 52(1):111–127.

Lee D, Hosanagar K, Nair HS (2018) Advertising content and consumer engagement on social media: evidence from facebook. *Management Science* .

Lehman DW, Kovács B, Carroll GR (2014) Conflicting social codes and organizations: Hygiene and authenticity in consumer evaluations of restaurants. *Management Science* 60(10):2602–2617.

Li X, Hitt LM (2008) Self-selection and information role of online product reviews. *Information Systems Research* 19(4):456–474.

Liu Y (2006) Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of marketing* 70(3):74–89.

Luca M (2016) Reviews, reputation, and revenue: The case of yelp. com. *Working Paper 12-016* Harvard Business School, Boston.

Luca M, Smith J (2013) Salience in quality disclosure: evidence from the us news college rankings. *Journal of Economics & Management Strategy* 22(1):58–77.

Mandler G (1982) The structure of value: Accounting for taste. *Center for Human Information Processing Report* 101.

Maslowska E, Malthouse EC, Bernritter SF (2017) Too good to be true: the role of online reviews' features in probability to buy. *International Journal of Advertising* 36(1):142–163.

McFadden D (1973) *Conditional logit analysis of qualitative choice behavior* (Frontiers of Econometrics, New York: Academic Press.).

McFadden D (1986) The choice theory approach to market research. *Marketing science* 5(4):275–297.

Mckinsey (2010) A new way to measure word-of-mouth marketing .

Mudambi SM, Schuff D (2010) Research note: What makes a helpful online review? a study of customer reviews on amazon. com. *MIS quarterly* 185–200.

Nelson P (1974) Advertising as information. *Journal of political economy* 82(4):729–754.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. *Marketing Science* 31(3):521–543.

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 79–86 (Association for Computational Linguistics).

Petrin A (2002) Quantifying the benefits of new products: The case of the minivan. *Journal of political Economy* 110(4):705–729.

Porter MF (1980) An algorithm for suffix stripping. *Program* 14(3):130–137.

Randall T, Terwiesch C, Ulrich KT (2007) Research note—user design of customized products. *Marketing Science* 26(2):268–280.

Rindova VP, Petkova AP (2007) When is a new thing a good thing? technological change, product form design, and perceptions of value for product innovations. *Organization Science* 18(2):217–232.

Sloman SA (1996) The empirical case for two systems of reasoning. *Psychological bulletin* 119(1):3.

Soll JB, Larrick RP (2009) Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35(3):780.

Stanovich KE, West RF (2000) Individual differences in reasoning: Implications for the rationality debate? *Behavioral and brain sciences* 23(5):645–665.

Sudhir K (2001) Competitive pricing behavior in the auto market: A structural analysis. *Marketing Science* 20(1):42–60.

Terwiesch C, Loch CH (2004) Collaborative prototyping and the pricing of custom-designed products. *Management Science* 50(2):145–158.

Train K (1986) *Qualitative choice analysis: Theory, econometrics, and an application to automobile demand*, volume 10 (MIT press).

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *science* 185(4157):1124–1131.

Ulrich KT, Ellison DJ (1999) Holistic customer requirements and the design-select decision. *Management Science* 45(5):641–658.

Yaniv I (2004) Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes* 93(1):1–13.

Yu M, Debo L, Kapuscinski R (2016) Strategic waiting for consumer-generated quality information: Dynamic pricing of new experience goods. *Management Science* 62(2):410–435.