Vertical versus Horizontal Variance in Online Reviews and Their Impact on Demand

Nah Lee, Bryan Bollinger, Richard Staelin

April 27, 2022

Abstract

This paper examines the differential impact of variances in the quality and taste comments found in online customer reviews on firm sales. Using an analytic model, we show that although increased variance in consumer reviews about taste mismatch normally decreases subsequent demand, it can increase demand when mean ratings are low and/or quality variance is high. In contrast, increased variance in quality always decreases subsequent demand, although this effect is moderated by the amount of variance in tastes. Since these theoretical demand effects are predicated on the assumption that consumers can differentiate between the two sources of variation in ratings, we conduct a survey that demonstrates that subjects are indeed able to reliably distinguish quality from taste evaluations from two subsets of reviews of size 5,000 taken from our larger datasets of reviews for 4,305 restaurants and 3,460 hotels. We use these responses to construct sets of reviews that we use in a controlled laboratory experiment on restaurant choice, finding strong support for our theoretical predictions. These responses are also used to train classifiers using a bag-of-words model to predict the degree to which each review in the larger datasets relates to quality and/or taste allowing us to estimate the two types of review variances. Finally, we estimate the effects of these variances in overall ratings on establishment sales, again finding support for our theoretical results.

Keywords: review variance, vertical and horizontal content, text analysis, machine learning, quality and taste variance, crowd-sourced data

INTRODUCTION

Consumers have long sought the opinions of others before deciding to purchase a product offering. Often these opinions focus on the latent quality level and the positioning of the offering. Prior to the availability of online reviews, most of this information was obtained in the form of word of mouth in personal conversations. However, now consumers often supplement this interpersonal information with online reviews that contain an overall evaluation of the reviewer's experience, often in terms of a star rating, along with an accompanying discussion of this experience. This paper focuses on how future potential consumers use not only the star ratings, but also the quality and taste variance inferred from the text content of the reviews to determine if they want to purchase the focal product offering. We center our attention on the service industry where customer experiences vary not only on the degree to which the service encounter meets the person's taste, but also on the quality of the delivered service. We show that potential customers can use the text discussions in the reviews to determine the source of the variance in the ratings. Specifically, they can assess a) the expected level of the quality of the service and the possible range of service encounters, and b) the positioning of the focal firm in terms of the specific features relative to the individual's ideal preferences, and the importance of the service not providing these ideal features. We then show how this acquired information affects future firm demand.

Our multi-method paper extends the analytic work of Sun (2012) and Zimmermann et al. (2018) by allowing the rating variance to be composed of two different, but continuous random variables, one associated with variation in the observed quality of the delivered service and the other associated with consumers' preferences with respect to the horizontal attributes of the service. In addition, we extend and the empirical work of X. Liu, D. Lee, and Srinivasan (2019), who look at the effects of both the ratings and the content of reviews on future demand (but does not explicitly consider the two types of heterogeneity, i.e., quality of service and preference in taste). Our focus is on industries where the rating variance comes from two very different sources that have very different implications. We make explicit why, and under what conditions, variances in prior experiences in quality may have a different impact on future sales than variances in expressed preferences for the taste aspects of the

service encounter. After deriving two implications flowing from our analytic model, we test specific aspects of the model using two surveys, one lab experiment, and two field studies. In each case we find strong support for the tested components and implications of the analytic model.

A key insight flowing from the analytic model is that the effects of changes in both types of variances depend on the level of the other variance. Thus, although increases in quality variance always decreases future sales, this effect is moderated by the amount of taste variance. Likewise, quality variance positively moderates the effect of taste variance on demand (while the mean rating negatively moderates it). Here, increases in taste variance increase future sales if and only if mean ratings are low and/or quality variance is high. Otherwise the effect is negative. The possible increase in sales is due to the fact that increases in taste variance imply higher average quality of the service as well as higher taste mismatch cost, holding fixed the mean rating. In other words, taste variance also signals the latent average quality of the establishment. One implication of this inter-relationship of the two variances is that when a firm invests in increasing its service reliability, it may find this strategy having only marginal value because this lower quality variance can change the effect of taste variance on future demand from being positive to being negative, a finding we find empirically.

The analytic model explicitly outlines a process by which readers determine the variances of the two types of review content, since this information allows second period consumers to learn about the firm's latent mean quality, service reliability (variance in quality), the firm's positioning and the importance of not having the firm provide the individual's ideal features (taste mismatch cost). Underlying this process is the assumption that readers of the reviews can partition the content of the reviews into quality comments and taste comments. Using survey data of 5,000 restaurant and 5,000 hotel reviews, we show consumers can reliably partition the content of these reviews into the two components. We then use this classification information to further predict the proportion of quality and taste content in each of over 900,000 text reviews for restaurants and hotels. Then, using the same assumed process that consumers use, we calculate the quality and taste variances for each restaurant and hotel in our sample. Next, using a controlled laboratory setting, we use a 2x2x2 within-

person design to show that consumers behave consistently with the predictions of our model, i.e., a) subjects have a lower purchase intent for an establishment that has a higher variance in the content discussing quality than one with a lower variance in quality-related content, all else equal, and b) subjects have a higher purchase intent for an establishment with a higher variation in the review content discussing taste issues compared to one with a lower variation in these reviews, all else equal, but only when the mean rating is low and/or the quality variance is high. Using the reviews and sales figures for restaurants in the San Francisco area and hotels located in Texas, we replicate the results using instrumental variable regressions, using the within-establishment variation in demand and ratings.

Our contributions are threefold. First, by building on the model of Sun (2012), we not only allow for the star ratings to vary with the two different product attributes, i.e., stochastic service quality and a mismatch of the individuals' preferences with the establishment's features, but also explicate the process by which consumers of a product offering write specific types of content in their reviews and how readers of these reviews use this information to update their prior beliefs about the mean quality level and its reliability, the positioning of the focal firm and the mismatch costs, all from the mean rating and the two variance components. This theoretical framework enables us to derive results regarding how vertical quality and horizontal taste attributes differentially affect sales outcomes. Importantly, these results have managerial implications for the effectiveness of service quality initiatives and repositioning strategies. Second, since these theoretical results assume consumers can reliably distinguish between vertically and horizontally oriented comments in reviews, we empirically show this ability in two different industry settings. Third, we show the demand predictions that result from our analytic model hold both in a controlled laboratory setting and in more generalizable field settings, and that these predictions differ from those in the prior literature. Finally, in the process, we illustrate that a simple and efficient machine learning methodology can be used to disentangle different sources of variance in reviews.

RELATED LITERATURE

Our investigation builds on a diverse set of literatures and topics including service quality and its effects on firm and consumer behavior, the effects of variable messages on consumer choice, and the effect of review variability on subsequent purchase behavior. One way of partitioning this literature is by whether the consumer has multiple experiences, and thus is trying to learn about the latent value underlying the experience, or is looking for *a priori* information to assess the expected latent value for the initial product experience. It is this latter situation that is most relevant for our study, since our second period consumers use the reviews to assess the latent quality value and the possible range of outcomes of their first encounter. However, findings pertaining to the former situation are relevant to the firm, since any firm behavior could affect returning customers (retention), possible future reviews from the new customers and thus the dynamic nature of reviews. With this noted, the dominant view is that consumers, in general, downgrade services that exhibit service quality variation both in terms of adoption (Meyer, 1981) and retention (Boulding et al., 1993; Rust et al., 1999; Sriram, Chintagunta, and Manchanda, 2015).

Although theoretically consumers do not like variation in service quality, they may respond positively to uncertainty in terms of ratings (but not quality). West and Broniarczyk (1998) propose that consumers form an aspiration rating level as their reference point, and their reaction to review dispersion is dependent upon whether the average rating is above or below the reference point: consumers dislike dispersion in ratings for products with the average rating above their reference point, but prefer review disagreement when it is below the reference point. Sun (2012) finds similar results, in which the effect of review variance associated with horizontal attributes depends on the average rating, although her results are derived, not from prospect theory, but in a game theoretic setting with risk-neutral consumers. In a similar spirit, Clemons, Gao, and Hitt (2006) provide empirical evidence that high variance items can serve as a hyper-differentiation strategy, which is helpful for brand growth and new product introduction. More recently, Rozenkrants, Wheeler, and Shiv (2017) find that people view polarizing products as a vehicle for self-expression, and prefer them when they experience low self-concept clarity and the product attributes are related

to self-expression (i.e., style but not quality).

The conflicting results in the literature about the effect of review variability on demand may stem from not specifying the source of variability in reviews. A large set of prior work has focused on products with deterministic quality, such as books (P.-Y. Chen, Wu, and Yoon, 2004; Chevalier and Mayzlin, 2006; Sun, 2012), movies (Y. Liu, 2006; X. Zhang and Dellarocas, 2006; Duan, Gu, and Whinston, 2008a; Duan, Gu, and Whinston, 2008b), and video games (Zhu and X. Zhang, 2010). For these products, most of the review variance likely stems from individuals' different preferences (although there certainly could be different opinions about objective quality of the product). On the other hand, service product offerings in industries such as restaurants, cruise ship vacations, airlines, hairdressers, spas and hotels naturally have variation in delivered quality, due to differing human and/or product interactions with each transaction. In these latter service situations, it is important to consider, not only review variance, but also the source of the review variance when examining its impact on future sales. In prior work on the impact of reviews on restaurants (Anderson and Magruder, 2012; Luca, 2016) and hotels (Vermeulen and Seegers, 2009), the review variance is ignored altogether. Only Sun (2012) considers review variance, in her case coming from heterogeneous tastes. No research, of which we are aware, has empirically addressed the differential effect of vertical and horizontal information variation on subsequent purchases.

With this noted, Zimmermann et al. (2018) most closely relate to our conceptualization, since they incorporate vertical quality risk into Sun (2012)'s theoretical framework. In contrast to our findings, they find that a higher variance caused by taste differences always results in a higher price and lower demand; but as they point out, their model varies from that of Sun (2012) in that the support for their mean rating has a higher lower bound than in Sun (2012), whose results depend on the mean rating being sufficiently low. Importantly, their model assumes that quality risk only manifests in terms of a possible quality failure, resulting in a rating of zero. This simplification makes their model parsimonious, while still incorporating variance coming from the vertical quality dimension. However, it also means consumers can easily differentiate between quality variance and taste variance, since quality failures are visible as zero ratings from the entire distribution of ratings. This sidesteps the question of whether consumers can disentangle the variance from the two sources in a

more general setting. Quality issues in many industries with stochastic quality realizations normally do not arise in a binary fashion, but instead occur to a varying degree for each consumption activity. Under this alternative data generating process, the quality variance would be blended with the taste variance in a way that the two cannot be distinguished just by looking at the distribution of ratings. Whether consumers can separate the two using review content is an empirical question, and one addressed in this paper.

Although they do not study review variance from taste-related reviews, Tucker and J. Zhang (2011) examine the interaction between "ratings" (actually popularity) and the breath of the market. They propose that low popularity may be due either to lower quality or narrower appeal. They note that "the same level of popularity implies higher quality for narrow-appeal products than for broad-appeal products." Similarly, we show that higher variance in reviews due to taste heterogeneity implies a higher mean quality level. In a different setting (i.e., earnings report), Harbaugh, Maxwell, and Shue (2016) use a Bayesian updating formulation to study the effects of ratings that come from multiple reports, on the reader's posterior beliefs and find empirical support that good news (i.e., high ratings) is more persuasive when the ratings are more consistent, and bad news (i.e., low ratings) is less damaging when the ratings are less consistent.

Finally, there are a number of recent papers (e.g., Hu, Pavlou, and J. J. Zhang (2017), Bondi (2019), and P. Chen et al. (2021)) that look at the dynamic aspect of reviews on the establishment's future customer base. In this new stream of research, the current reviews affect which consumers choose to purchase the product today, possibly changing the composition of subsequent reviewers. Such dynamics may lead to cyclical patterns in ratings and demand, in which a lower rating leads consumers to buy the product only if there is a very good taste match between them and the product, which then leads to higher ratings, which then leads to higher demand from consumers with lower taste match, who then leave lower reviews, and so on. Although we do not directly model such dynamics, we briefly discuss the implication of our empirical results to review dynamics in the conclusion section.

MODEL

We develop a model that explicates the process by which first period customers write reviews of their service experience and second period potential customers use this review information to determine if they want to choose this focal establishment. In this way, we derive the firm's second period demand function which is a function of our three key review variables, namely the mean review rating (M) and the two variance components of this rating, one associated with the firm delivering stochastic quality experiences (V^q) and the second coming from customers having heterogeneous preferences for the firm's horizontal features (V^t). Then using this second period demand, we derive comparative statics to generate our two testable hypotheses.

Given our interest in the service industry we broaden Sun (2012)'s model (which assumes ratings vary only due to heterogeneous consumer preferences), by allowing the ratings to also vary due to variances in the firm's delivered service. Thus, the product in our model (a good or service) is characterized along two dimensions, 1) a quality distribution from which stochastic quality is drawn, due to factors such as server variability (vertical dimension), and 2) the product's positioning relative to individuals' preferences for its features (horizontal dimension). Following the convention of the service quality literature, e.g., Boulding et al. (1993) and Rust et al. (1999), realized quality is a random variable. This has two implications. First, the uncertain service quality implies that consumers face an *a priori* risk. Second, since most consumers do not like uncertainty, i.e., they are risk adverse, quality variance in the service industry is bad. Moreover, since quality lies in the vertical dimension, higher quality results in higher utility for all consumers. Holding quality fixed, different consumers can also experience different utility for an offering due to the firm's positioning not meeting the specific individual's ideal preference. We model heterogeneity only in consumer tastes, assuming identical consumer valuations for quality. We model taste preference as a fixed, time-unchanging, idiosyncratic attribute for each consumer for each product. In our model, the uncertainty in the taste mismatch gets resolved through reviews, while the uncertainty about quality is quantified in terms of the average quality level and its variance.

A key feature of our model is the explication of a process in which first period customers

decide on what information to convey in their reviews and how second period readers interpret and use this information. Specifically, we assume a) these first period consumers write about experiences that deviate from their expectations, and b) the percentage of the review devoted to quality (q_i) versus taste ($1 - q_i$) reflects the relative magnitude of the deviations between 1) the realized and expected quality, and 2) the individual's taste mismatch versus the average consumer's mismatch.[1] The second period consumers know this is the process that generates the reviews, and thus they can partition the reviewer's deviation from the mean rating for all of the focal firm's reviews into the quality deviation and the taste deviation. Once these two deviations are known for all of the focal establishment's reviews, the second period consumer can calculate the two component variances, V^q and V^t, which allows the individual to determine, not only the establishment's average quality level and reliability of the service, but also the positioning of the firm as well as the individual's mismatch costs. Consumers then use this information to decide whether or not to buy the product after the firm sets the price for the second period. Although our primary model specification assumes firms adjust prices (optimally) in the second period, as in Zimmermann et al. (2018), our results are robust to an alternative scenario in which prices are held constant or set at some other observed level in the second period.[2]

Given this overview, we next lay out the specific elements of our model. We characterize a product experience in terms of quality and taste mismatch cost. Quality, which is stochastic in nature, is uniformly distributed from $\bar{v} - r$ to $\bar{v}$, where $\bar{v}$ is the maximum possible delivered quality and r is the range of possible delivered quality experiences. Consumer tastes are uniformly distributed in the horizontal dimension. Consumers will consider purchasing a product if they lie within a distance of 1 from a particular product's taste location. Consumers are fully aware of their ideal taste preference, i.e., their exact location in the taste dimension. When a consumer located at $x \in [0, 1]$ distance away from the focal product purchases it at a price p and consumes it, his/her utility is $\dot{v} - t \cdot x - p$, where $\dot{v}$ is

[1] This assumption is similar to the Bayesian concept that data only has value if it alters a person's prior belief. It also has behavioral support in that consumers like to talk about experiences that "excite" or "frustrate" them.

[2] In fact, we control the second period price in our laboratory experiment, in contrast to our field studies, where we assume that sellers provide appropriate price response.

the realization of stochastic quality and $t > 0$ is the taste mismatch parameter.[3]

Before the anticipated product launch, the product's quality and taste mismatch cost are unknown to both the seller and the consumers. However, everyone has unbiased prior beliefs for $\bar{v}$ and r, and the joint probability density function is denoted $f(\bar{v}, r)$. Likewise, it is common knowledge that each product has a fixed attribute t, and the unbiased prior belief on t is $g(t)$.

Given this uncertainty, the indifferent first period consumer is located at D_1 distance away from the product, such that

$$E_{\bar{v},r,t}(\textbf{C.E.}[v - t \cdot D_1 | f(\bar{v}, r), g(t)]) - p_1 = 0, \tag{1}$$

where $\textbf{C.E.}[v - t \cdot x]$ stands for the certainty equivalent of the utility derived from uncertain product attributes, for a consumer at x distance from the product. Assuming identical prior beliefs and degree of risk aversion, any consumers located at $x \in [0, D_1]$ from the product would purchase it, making the first period demand D_1, with a unit mass assumption.[4]

Once each consumer visits the service, v and t are realized and (by assumption) each consumer leaves a product rating reflecting his or her realized utility: $s(x) = \dot{v} - t \cdot x$. The distribution of these ratings reflects the distribution of the consumer utilities, shifted by the price, and as shown in Figure 1, is the sum of two uniform distributions.[5,6]

It is easy to show that the mean (M) and the variance (V) of these ratings left by the first period consumers are:

$$M = \bar{v} - \frac{r}{2} - \frac{t \cdot D_1}{2} \qquad \text{and} \qquad V = \frac{1}{12}(r^2 + t^2 D_1^2), \tag{2}$$

[3] By assuming that consumers within a distance of 1 from a product are the only ones to consider it, we are normalizing on x across the products and have t capture the variance in the taste dimension for each product.

[4] A necessary condition for sales to occur in the first period is $E_{\bar{v},r}(\textbf{C.E.}[v | f(\bar{v}, r)]) - p_1 > 0$. A consumer whose taste is exactly matched with the product ($x = 0$) would purchase it.

[5] We do not consider binning or censoring for the ratings, so theoretically the domain for ratings is $[-\infty, \infty]$.

[6] We admit our two-stage model is a discrete simplification of what in reality would result in is a continuous evolvement of review distributions. However, even if we abstract away from the uniform distributional assumptions for both quality and taste space, and thus the mean and the variance no longer contain full information about both product attributes, we find that once consumers can correctly distinguish the two underlying distributions of quality and taste, the qualitative insights driven from our simpler model continue to hold as long as consumers are risk averse towards the stochasticity in quality, and truth telling is satisfied so that consumers can correctly identify the taste parameter of each product. This is why it is important for consumers to be able to distinguish the two dimensions of product attributes from others' feedback.

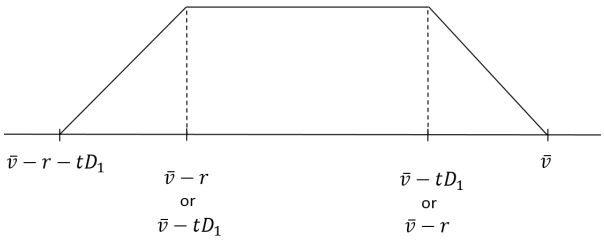

Figure 1: Early Stage Distribution of Ratings

in which D_1 is the first period demand.[7]

Note that the variance of the ratings across these first period individuals can be decomposed into two parts: the part associated with the product's delivered stochastic quality, as captured by the range of possible quality outcomes, r, and the part associated with the taste mismatch parameter, t, i.e.,

$$V^q = \frac{1}{12}r^2 \qquad \text{and} \qquad V^t = \frac{1}{12}t^2 D_1^2. \tag{3}$$

These variances reflect the variations in the individuals' quality and taste deviations from their expectations, which according to our assumed review writing process determines the proportion of their reviews devoted to quality and taste. For the remainder of the paper, we let V^q denote the partial variance arising from quality variation and let V^t denote the counter-part variance arising from taste mismatch, where the sum of the two equals the total variance, V, i.e., $V = V^q + V^t$.

We next assume the second period consumers, after reading the first period reviews, are capable of decomposing each rating into these two components, in expectation.[8] Once they have determined the quality and taste deviations across many reviews, the second period consumers can infer V^q and V^t. Using this information, along with the observable mean rating M (and the first period demand D_1, which can be inferred[9]), the second period consumers can determine the true underlying product attributes, i.e., the quality distribution parameters, $\bar{v}$ and r, and the taste mismatch parameter, t, and thus the disutility of not

[7] Our model accounts for risk aversion due to uncertainty in the $\bar{v}$, r, and t as well, which will impact the first period demand and through this demand, the distribution of ratings. With more uncertainty in the prior beliefs of $\bar{v}$, r, and t, only consumers with less of a taste mismatch will purchase in period 1.

[8] The detailed discussion of how consumers infer the expected V^q and V^t is found in the Appendix.

[9] The first period demand D_1 can be inferred by the second period consumers if they share the same prior beliefs and risk aversion with first period consumers and know the first period price, i.e., the second period consumers derive D_1 from equation (1): $\int\int \mathbf{C.E.}[v - t \cdot D_1 | f(\bar{v}, r), g(t)] \cdot dF(\bar{v}, r) dG(t) = p_1$.

getting the most preferred option, i.e.,

$$\bar{v} = M + \sqrt{3V^q} + \sqrt{3V^t}, \qquad r = 2\sqrt{3V^q} \qquad \text{and} \quad t = \frac{2\sqrt{3V^t}}{D_1}. \tag{4}$$

Based on this knowledge, the second period consumers fully resolve any uncertainty in t, and can make their purchase decisions, conditional on second period prices. This allows us to determine second period demand, i.e., D_2 is derived by noting that the indifferent consumer at the second period satisfies

$$\textbf{C.E.}[v|\bar{v}, r] - t \cdot D_2 - p_2 = 0 \tag{5}$$

where $\textbf{C.E.}[v|\bar{v}, r]$ depends on the now-known $\bar{v}$ and r.[10,11,12] Expecting this demand function, the seller finds the optimal second period price that maximizes his/her profit by solving: $\max_{p_2} \; p_2 \cdot (\textbf{C.E.}[v|\bar{v}, r] - p_2)/t$. The equilibrium demand for the second period is found as:

$$D_2^* = \frac{\textbf{C.E.}[v|\bar{v}, r]}{2t}. \tag{6}$$

To further solve for (6), we assume constant absolute risk aversion (CARA), i.e., the degree of risk aversion is constant. CARA utility is represented as $U(v) = 1 - e^{-\alpha v}$, with the absolute risk aversion coefficient $A(v) = -\frac{U''(v)}{U'(v)} = \alpha > 0$. The certainty equivalent of v is then derived from the definition of certainty equivalent leading to the following:

$$\textbf{C.E.}[v|\bar{v}, r] = \bar{v} - \frac{1}{\alpha}\left(ln\frac{1}{\alpha r} + ln(e^{\alpha r} - 1)\right). \tag{7}$$

Substituting $\bar{v}$ and r from equation (4), we get

$$\textbf{C.E.}[v|M, V^q, V^t] = M + \sqrt{3V^q} + \sqrt{3V^t} - \frac{1}{\alpha}\left(ln\frac{1}{2\alpha\sqrt{3V^q}} + ln(e^{2\alpha\sqrt{3V^q}} - 1)\right). \tag{8}$$

Substituting (4) and (8) into (6), we find the second period equilibrium solution for demand in terms of M, V^q, and V^t:

$$D_2^* = \frac{D_1}{4} \cdot \frac{M + \sqrt{3V^q} + \sqrt{3V^t} - \frac{1}{\alpha}\left(ln\frac{1}{2\alpha\sqrt{3V^q}} + ln(e^{2\alpha\sqrt{3V^q}} - 1)\right)}{\sqrt{3V^t}}. \tag{9}$$

Using this demand function, we can determine the marginal effect of the three observable

[10] A necessary condition for sales to occur in the second period is $\textbf{C.E.}[v|\bar{v}, r] - p_2 > 0$.

[11] We assume that market is never fully covered to avoid discussion of corner solutions. A sufficient condition for incomplete market coverage is: for products with stochastic quality $v \in [\bar{v} - r, \bar{v}]$, $\max(\bar{v}) < \min(t)$.

[12] See Online Appendix OA for a summary of information sets in the consumer decision process.

review statistics, M, V^q, and V^t, on the second period demand:

$$\frac{\partial D_2^*}{\partial M} > 0, \ \frac{\partial D_2^*}{\partial V^q} < 0, \text{ and } \frac{\partial D_2^*}{\partial V^t} > 0 \text{ iff } M < -\sqrt{3V^q} + \frac{1}{\alpha}\left(ln\frac{1}{2\alpha\sqrt{3V^q}} + ln(e^{2\alpha\sqrt{3V^q}} - 1)\right). \tag{10}$$

Moreover, the conditional effect of V^t implies $\frac{\partial^2 D_2^*}{\partial V^t \partial M} < 0$ and $\frac{\partial^2 D_2^*}{\partial V^t \partial V^q} > 0$.

These comparative static results lead to the following predictions on how the second period demand is altered by each type of variance:

Hypothesis 1. An increase in review variance due to quality inconsistency across reviewers' experiences always leads to lower sales.

Hypothesis 2. An increase in review variance due to taste mismatch costs leads to higher sales if and only if the mean rating is sufficiently low and/or the quality variance is sufficiently high; otherwise, taste variance leads to lower sales.

H1 is intuitive and is addressed in Zimmermann et al. (2018) and Boulding et al. (1993), albeit with different model assumptions. In addition, since demand is a multiplicative function of V^q and V^t, we expect the marginal effect of V^q to depend on the level of V^t. We discuss this moderating effect in more detail subsequently. H2 reconciles the contradictory results found in Zimmermann et al. (2018) and Sun (2012) by noting how the actual sign reversal of V^t depends on not only M as in Sun (2012), but also V^q for products with stochastic quality. This sign reversal is made clear in Figure 2 where we plot the second period demand against the taste variance, V^t, and a composite index, w (which increases in the mean, M, and decreases in the quality variance, V^q). V^t positively affects second period demand if and only if w is less than zero.

Before discussing the intuition behind H2, we note that the two hypotheses are "all else" statements, since they come from comparative statics analyses. Consequently, they can be viewed (and tested) by comparing two firms who differ only on one particular review variable at one point in time (as in our laboratory experiment) or by observing how within-firm changes over time in a given review variable affect sales (as in our field studies).

With this as a preamble, we now discuss the intuition behind the effect of V^t in H2. Holding fixed M and V^q, consumers can determine that a firm with a higher V^t in ratings,

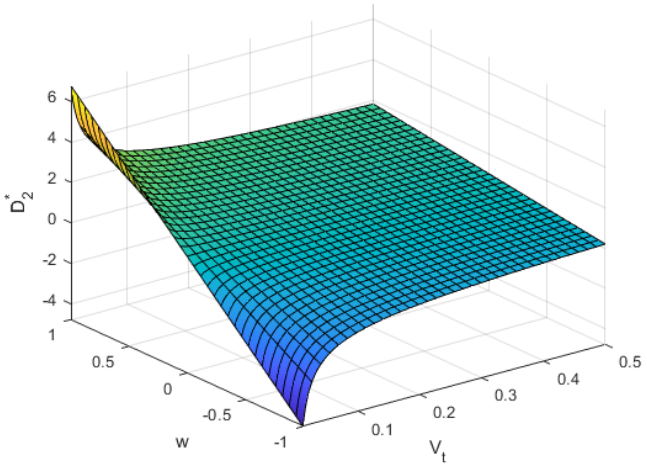

Figure 2: Second Period Demand; plotted against V^t and $w = M + \sqrt{3V^q} - \frac{1}{\alpha}\left(ln\frac{1}{2\alpha\sqrt{3V^q}} + ln(e^{2\alpha\sqrt{3V^q}} - 1)\right)$ (a constant $D_1 = 4$ is assumed)

compared to an otherwise identical firm, indicates that the taste parameter t is large (i.e., they should seriously consider the cost of taste mismatch). However, as seen from equation 4, it also means that the underlying average quality level for the focal firm, $\bar{v} - r/2$, is higher than the otherwise identical firm. This results in two counter-balancing effects: second period demand shifts outward due to high average quality, increasing the number of potential customers, but the high cost of mismatch implies that fewer second period customers would choose the focal product offering. Which effect is greater depends on the quality variance, V^q, as well as the mean of the first period rating, M. If M is low (and/or V^q is high), then a larger V^t results in larger second period sales (see Figure 2). In this situation, the effect of higher average quality level is greater than the effect of the taste mismatch cost. However, if M is high (and/or V^q is low), then the taste mismatch effects is greater, and a larger V^t results in smaller second period sales for the focal firm.

Although these implications provide very specific predictions on how demand responds to changes in the statistics of the review ratings, H1 and H2 only follow if consumers can distinguish between the two different sources of variance and use this information during their purchase decision process. To date, the differential effects of the two different sources of variance have never been empirically tested. Therefore, we next demonstrate that consumers are able to disentangle the two sources of rating variance. Then in subsequent sections, we test our two demand hypotheses in a controlled laboratory setting and two field studies, one

for restaurants and the other for hotels.

PRODUCT EXPERIENCE DATA

Consumer Reviews

We start our discussion by describing the two large datasets (restaurants and hotels) that we use in our subsequent analyses. We focus our attention on industries where, *a priori*, we believed product experiences reflect both vertical and horizontal attributes. Restaurants and hotels fit this description well. This led us to collect reviews for businesses in these two industries that are found on "Google Places". Google allows business owners to register their firms and post information about the business. This information is displayed when consumers search on Google Maps, along with the reviews posted by their customers (see Figure 3). These reviews always include a star rating, and about 60% of the restaurant reviews sampled and 70% of the hotel reviews sampled also include text detailing the customer's experience. Once these establishments were matched with revenue data, the vast majority of matched businesses (over 95%) were found to have text reviews. We define our estimation samples as the two sets of matched data.

The first sample of businesses includes 4,305 restaurants, bars, cafes, and bakeries, the overwhelming majority being restaurants, located in San Francisco, CA and neighboring cities; the second sample includes 3,460 franchised lodging establishments located in the State of Texas.[13] We collected the restaurant sample of reviews in May 2018. The first reviews were written in the early 1990's, although almost all of the reviews are post-2010, which is when the service was integrated into its current format.[14] The hotel reviews were collected in July 2020 and span the years 2006 to 2019, although most are post-2010.[15] When

[13]The exact target area for the restaurant sample includes the latitude range of [37.690, 37.906] and the longitude range of [-122.518, -122.200] in degrees. For our hotel sample, we limited our scope to franchised lodging establishments who file their monthly revenues to the State of Texas, in order to exclude Airbnb's and other individual vacation rentals, etc.

[14]A large number of reviews were posted in 2017 (due to an exponential growth Google Reviews experienced in 2016-2017) but the remaining reviews were approximately evenly distributed over the remaining six years.

[15]We dated reviews using Google's posted date for the review, e.g., 3 weeks ago, 1 year ago, etc. Thus, the posted times are approximations of the true time when the given review was available for a consumer to view it.

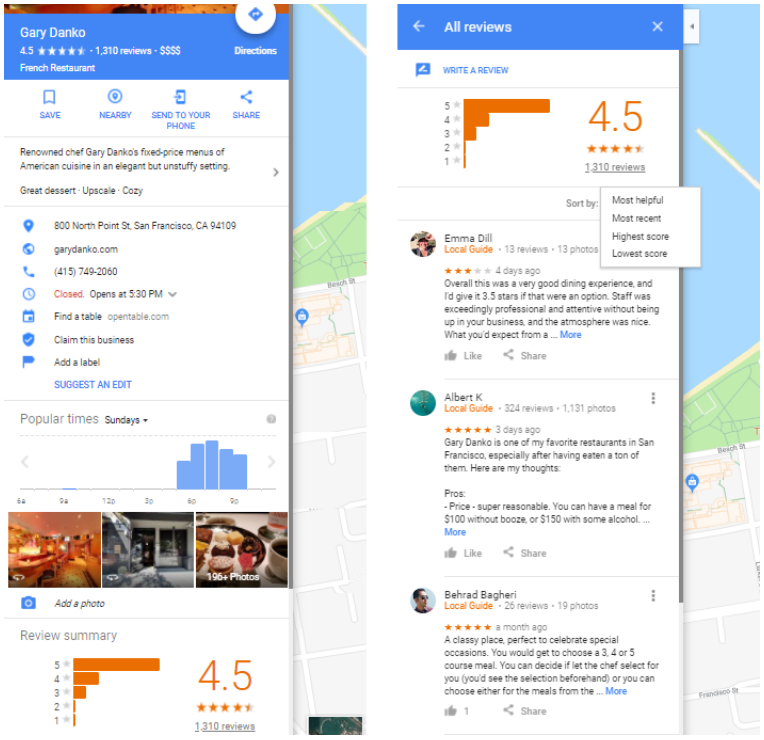

Figure 3: Business Information Landing Page (left) and Business Reviews (right) on Google

calculating review statistics for a given establishment at a given time point, we assumed that the reviews posted up to the date under investigation were available to consumers.

Given our setting, each rating incorporates both the quality level experienced by the reviewer and that person's taste mismatch cost. However, only if the rating is accompanied with review text can consumers disentangle the contributions of these two sources on the rating variation. Consequently, we limit our analysis to reviews with texts in the main analyses.[16]

[16]We compared the mean of the ratings given without any text to those with text. We find these means are very similar; the average across all restaurants (hotels) is 4.245 (4.079) for rating-only reviews and 4.184 (3.873) for text reviews, although the average variances are slightly lower for no-text reviews. Thus, there appears to be little concern for selection bias. In addition, when we include the statistics of the ratings with no text as control variables in a robustness check, their effect is insignificant and the effect of the text reviews remains the same.

Content Coding

Overview

The analysis plan for our samples of reviews centers on three different (yet related) objectives. First, we wanted to determine if consumers can reliably separate the review comments into statements on vertical quality and horizontal taste matches. Second, assuming the respondents were successful in identifying these two aspects of a given review, we wanted to use the information from this survey of respondents to better understand which words are associated with statements about quality versus taste for the focal industry. This understanding of word use allowed us to create reviews in which we manipulated M, V^q, and V^t within a laboratory experiment. Third, we use these responses to train two classifier algorithms for each sample that calculate the degrees to which each review discusses quality-, (i.e., our q_i measure), and taste-related topics $(1 - q_i)$, based on the text of the review. We used this classification for each review, along with the deviation of the reviewer's rating from the mean rating, to calculate an expected V^q and V^t for each establishment in our total sample in our field studies. We describe our text analysis methodology and the responses of the classification surveys next.

Methodology

We started with the entire set of 283,069 restaurant (634,121 hotel) text reviews, which contained 28,572 (31,080) unique words in total. We reduced the dimensionality by eliminating words that occurred only once, since their effect on classification would be minimal. This left us with 17,902 (18,577) words for restaurants (hotels). We selected a sample of 5,000 text reviews for both industries to be "hand" classified by survey respondents. Our goal was to choose the 5,000 restaurant (hotel) reviews that offered as large a coverage of the words as possible, favoring "important" words, i.e., ones that affect the ratings. Specifically, we ran elastic net regression (which combines L_1 and L_2 penalties) of ratings on the set of entire terms to identify words that were highly associated with the ratings. Iteratively using sets of regulation thresholds, we then approximately matched the number of surviving words to what 5,000 reviews could possibly contain. The net result was the identification of a set

of 5,000 restaurant (hotel) reviews that contained 12,798 or about 71% (10,211 or about 55%) of those 17,902 (18,577) words. We then assigned these sets of 5,000 reviews to two different groups of workers (one for restaurants, one for hotels) to classify their assigned set of reviews along the two dimensions of interest. We note in passing that these 5,000 reviews also contained 4,983 (6,114) words that only appeared once, so these 5,000 reviews contained 17,781 (16,325) words in total.

After obtaining this subset of 5,000 reviews for each industry, we divided them into bundles of ten reviews, resulting in 500 surveys. The surveys were distributed to experienced Amazon Mechanical Turk (AMT) workers who were asked to classify the parts of each review in their bundle associated with quality and with taste. The length of each review ranged from about ten to 250 words. In order to keep the task similar for each survey, we grouped the reviews keeping the total length of each set approximately equal. Hence, each survey of ten reviews consisted of three long, four medium length, and three short reviews, i.e., length (and not content) was the criterion used to place the reviews in batches of ten. Workers were asked to read the generic definitions of quality and taste statements, i.e., quality concerns aspects of the experience where everyone would agree high quality is better than low quality, and taste refers to aspects where individual preferences differ across the population. The workers were then asked to read the first text review in full and then pick out the phrases or sentences that are related to quality issues. Then, they repeated this process for personal taste or fit issues expressed in that first review. Phrases not selected were considered as unrelated to that topic and thus the rating. (see Figure 4 for an example response to these two tasks for restaurants).

These text selection questions enabled us to let the respondents decide how they distinguish vertically-oriented quality issues from horizontally-oriented personal taste issues. Some respondents selected an entire paragraph as a discussion about a topic, while others directly pointed to specific words or phrases that they felt were central to the discussion, thereby providing more focused measures of which words are driving the nature of the review. Since the workers were free to use any classification strategy, some sentences were selected as being related to both quality and taste, while at other times answers were left blank if the worker felt the review covered only one or neither dimension. Additionally, the workers

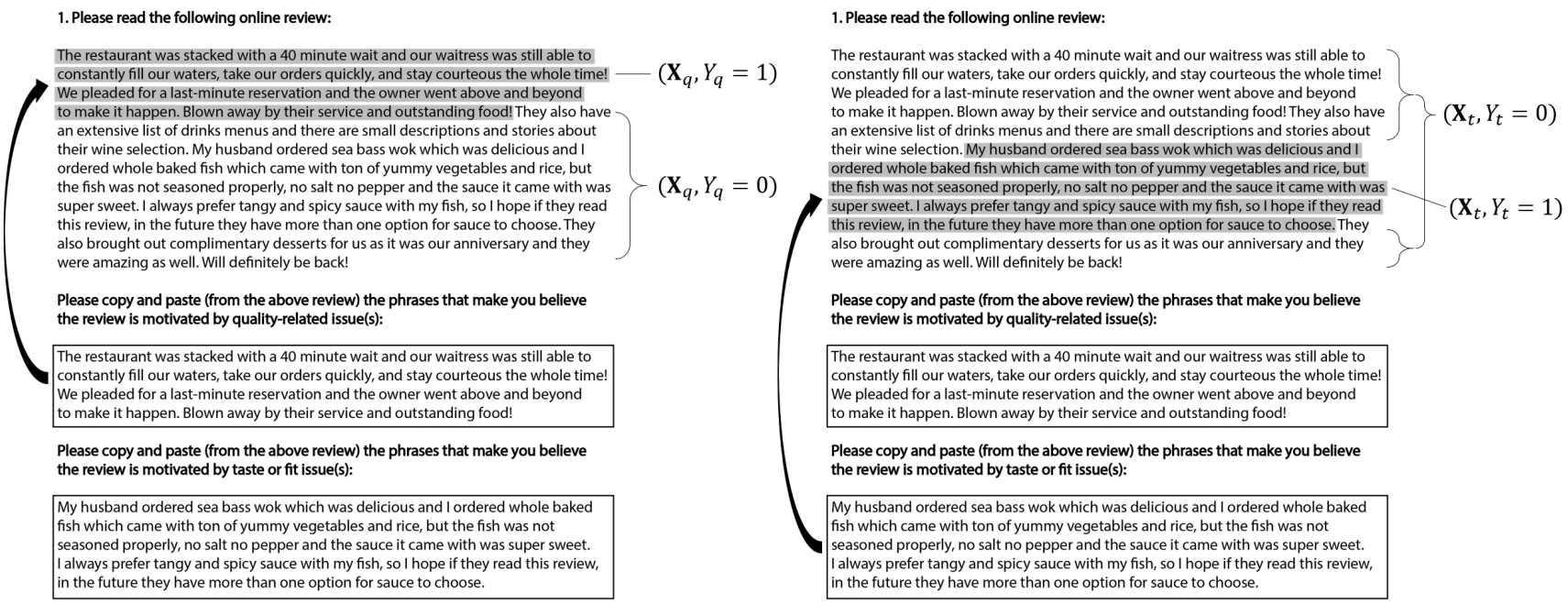

Figure 4: Text Selection Question Measuring Which Words are Related to Quality Issues (left); Personal Taste Issues (right). $\mathbf{X}_{\text{q}}$indicates a vector of predictors (i.e., words) for Y_q (i.e., whether or not a quality topic exists). Similarly, the variables for a taste topic are denoted with t subscripts.

were asked two multiple choice questions associated with each review, one indicating the relative importance of quality- versus personal taste-related issues and the other the overall helpfulness of the text review. The respondent repeated this process nine more times, once for each review.

We took two approaches to ensure that the responses were high-quality and reliable. First, we screened all answers by comparing an individual's multiple choice response to the the relative importance of quality- versus taste-related issues question with the amount of text input provided for quality- and taste-related content for a given review. We used this comparison to identify and filter out fraudulent submissions (those automatically filled out by bots, or someone who picked random or identical answers to all questions). We also included an attention check question near the end of the survey (see Online Appendix OB for the format of this checkpoint), to assess that the worker was reading each review carefully. Any submission that failed to answer this attention check correctly was dropped.

Second, we selected 150 reviews (representing 15 unique sets of reviews) to be analyzed by between 6 to 12 (4 to 6) unique respondents for each set of classifier reviews. We used these responses to calculate a Cronbach's Alpha for each set of reviews, based on the answers to the multiple choice questions. The average across 15 statistics for restaurants (hotels) is .824 (.900), while the median is .814 (.909). All sets had a Cronbach's Alpha greater than the commonly acceptable threshold of .7. Although the rest of the 485 surveys in each survey were rated only once, we note that many of the words (and especially the frequently occurring ones) appeared multiple times throughout the samples of 5,000 reviews. Hence, the words related to the common topics were actually rated many times. We provide the exact survey instrument and other details of the administration in Online Appendix OB.

Survey Responses

We next examined which words the AMT workers identified as being in the quality- and taste-related phrases and found that some words occur frequently in discussing both topics. Consequently, in order to determine which words are more strongly associated with one type of topic rather than the other and thus better for classifying a review with respect to its quality or taste focus, we calculated the Z-test statistics for differences in two population

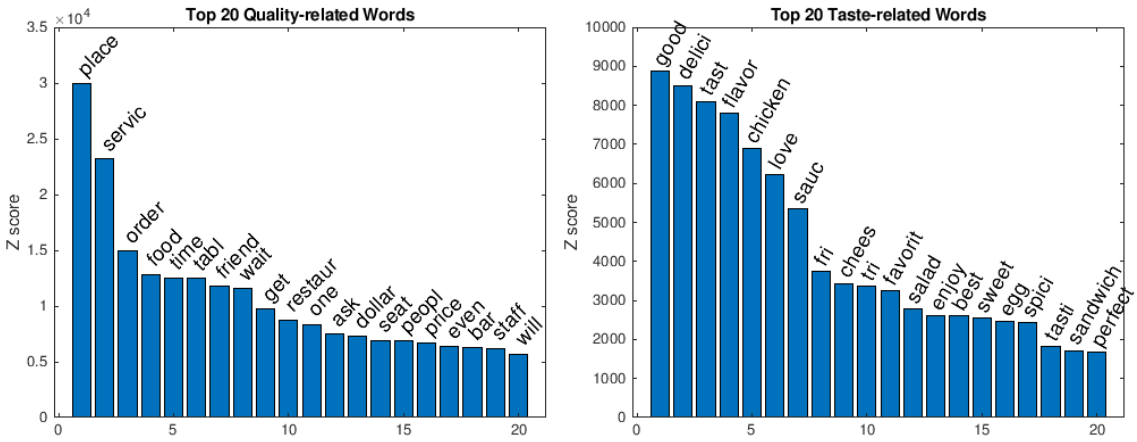

Figure 5: Words Most Likely to be Quality-related (left) and Taste-related (right) in Restaurant Reviews

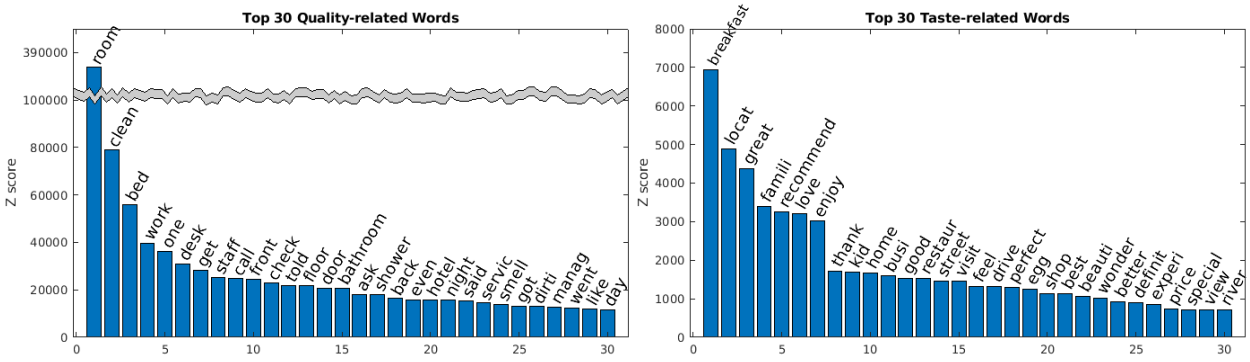

Figure 6: Words Most Likely to be Quality-related (left) and Taste-related (right) in Hotel Reviews

proportions.[17] Based on the statistic values, we were able to identify words most likely to be quality-related, rather than taste-related, and vice versa (see Figures 5 and 6). For example, the words service, order, wait, seat, staff, place, and price (including the "$" sign, which was coded as the word "dollar") are among those found in the quality-related comments for restaurants. On the other hand, specific menu items as well as the descriptive adjectives regarding food taste (e.g., sweet, spicy), ambience, location, parking, etc., are found to be more likely in personal taste-related comments. We used this information to help us create the forty restaurant reviews to use in the laboratory experiment, which is discussed next. In addition, the survey responses were also used to train classifier algorithms for predicting the

[17]The Z-statistic for testing for the difference in two population proportions is: $Z = (\hat{p_1} - \hat{p_2}) \,/\, \sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}$ where $\hat{p_1}$ and $\hat{p_2}$ are proportions, and n_1 and n_2 are the total counts in two populations, and $\hat{p}$ is the overall proportion in the entire population.

existence of quality and taste topics, and the details are discussed in the field study section.

LABORATORY EXPERIMENT–FOR RESTAURANTS

Overview of the Design

The goal of the laboratory experiment was to explicitly test our two hypotheses, which not only involve main effects but also interactions between our review variables M, V^q, and V^t. We do this via a 2 (low and high M) x 2 (low and high V^q) x 2 (low and high V^t) within subject design, in which each respondent looks at all eight restaurants (one per each cell) and the associated set of reviews. We vary M, V^q, and V^t by manipulating the star ratings and text of the reviews associated with a given restaurant. To keep the task manageable and to hold the number of reviews fixed, each of the eight restaurants had five unique reviews, each consisting of a star rating and a short text review about the reviewer's experience. In order to control for prior beliefs about the restaurants, we fixed the general location (local), cuisine (American), menu types, price, and review volume to be the same for all eight restaurants. Respondents were also told that these hypothetical restaurants were new to the respondent. Additionally, we controlled for the order of information presentation by randomizing the order of the restaurants across respondents, as well as the order of the five reviews associated with each restaurant. After reading the five reviews for a given restaurant, the respondent provided his or her intent to dine at that restaurant (our measure of sales).

This 2 x 2 x 2 design allowed us to compare restaurants that vary with only one variable of interest (M, V^q, or V^t), holding the other two variables fixed, in order to assess what happens if a restaurant changes only along that dimension. This allowed us to test H1 by determining if subjects' stated purchase likelihoods decrease with increases in V^q and test H2 by determining if purchase intent increases with increases in V^t from the base case (high M and low V^q) as M becomes low and/or V^q becomes high.[18]

[18]We also tested the significance of the effects in the end points, i.e., low M and high V^q vs. high M and low V^q, since our theory only predicts that the change in sales switches somewhere over the range of the values in the M and V^q pair. We specifically tested that the effect of V^t is significantly positive under "low M and high V^q" condition, while it is significantly negative under "high M and low V^q". Technically, we only needed to test if the effect in the former condition is greater than the effect in the latter condition.

Manipulated Reviews

We constructed 40 different reviews (5 reviews for each of the 8 restaurants) that were used to vary M, V^q, and V^t. The text of each review describes experiences that reflect vertical and/or horizontal features. By constructing different sets of reviews, we were able to manipulate the mean of the reviews, M, and the variances due to vertical and horizontal experiences. We focused the review content for a specific review on just quality, just taste, or equally on both. (For each restaurant, four of the reviews only discussed either quality or taste, i.e., q_i was 0 or 1, while the fifth described both dimensions equally, i.e., $q_i = .5$). The text was written to be compatible with the star rating given, where positive reviews were associated with 4- and 5-star ratings, negative reviews with 1- and 2-star ratings, and neutral reviews with a 3-star rating. For example, if the restaurant was associated with a high V^q, then the text across the five reviews for this restaurant would report a large dispersion of opinions on the quality of the restaurant. Relying on findings reported earlier on words most likely to be associated with quality (taste) issues, we constructed text reviews talking about wait times and interactions with the server in order to manipulate the low and high quality variance within the set of reviews, while reviews discussing the horizontal attribute of the spiciness of the menu items were used to manipulate V^t within a restaurant. For example, one text review focusing on the taste dimension reads: "The spicy chicken was too spicy for me. Too spicy that I couldn't finish my food and didn't enjoy it at all. Maybe others will like it. (2 stars)". Each review was written to be within a similar length, and the total amount of review text for each restaurant was also similar.

We assigned restaurants 1-8 to experimental cells 1-8. Figure 7 shows the exact star ratings we used for the reviews to manipulate the quality variance and the taste variance for the four (2 x 2) restaurants with a low average rating, which we designated as being in experimental cells 1-4. For example, for restaurant 1 (i.e., cell 1), there were 3 reviews that discussed quality issues and also indicated variation in star ratings for these reviews. This restaurant also had three reviews that mentioned taste issues, but there was no variation in star ratings for these three reviews. Thus, this cell 1 restaurant represents the high V^q, the low V^t cell. The four high-mean restaurants were in experimental cells 5-8 and used the same

Figure 7: Manipulation of Quality Variance and Taste Variance for Low Mean Restaurants

ordering as for cells 1-4; but now each rating was shifted by +1 star for all 20 ratings resulting in 20 new (more positive) reviews. Thus, the four restaurants with high mean ratings had a mean of 4 stars, while having the identical (total and partial) variance structure as the low mean restaurants. The 40 different written reviews used in the experiment are available in Online Appendix OC.

We pretested our manipulations of quality variance and taste variance using a sample of 45 AMT respondents who were asked to quantify the amount of quality variance and taste variance associated with each of the eight restaurants and who also passed a screening task. All 8 of the appropriate contrasts were statistically significant and in the correct direction. The pooled contrast of all pairs of comparisons, after each subject's responses were standardized, also shows significant differences ($p < .0001$) and in the desired direction. The survey instructions and questions and the standardized cell means for the pretest survey are available in Online Appendix OC.

Survey Questions and Administration

The experiment was administered in a laboratory setting. The respondents (average age of 29.6 years, ranging from 18 to 75; 63% female) were part of a southeastern university's Behavioral Research pool who have expressed interest in participating in research studies,

and in our case the respondents took part in multiple short studies in one sitting. After completing all the studies they were compensated for their participation.

At the beginning of our experiment, respondents were instructed that they were to decide how likely they were to go to a number of different local restaurants that were new to them. Then, the respondents were shown the five reviews, in a random order, for the first randomly presented restaurant on a computer screen. Below the reviews on the same screen page, the respondents indicated their purchase intent for the restaurant by responding to the question "How likely are you (personally) to choose this restaurant?" using the seven item scale anchored by "Not likely at all" and "Highly likely". This process was repeated seven more times on the following screen pages for the remaining seven restaurants. Respondents were encouraged to navigate back and forth between restaurants in order to scale their answers across restaurants. The exact survey instructions and instruments are available in Online Appendix OC.

After completing the survey questions for the eight restaurants, the respondents were presented with a ninth restaurant where one of the five new, but otherwise similar, reviews had a sentence embedded in the text instructing the respondent to choose a specific answer for the questions that followed. This task was given to verify that the person was paying strict attention to the task at the end of the study.

Experiment Results

Out of the 178 subjects who finished our survey, N=90 (50.56%) read each review carefully enough to pass our very subtle attention check. Given the low rate of "attention", we first compared the two groups and looked for differences between them. For the 90 subjects who passed the attention test, their average age was 29.9, 69% were female, and they took on average 286 seconds to complete the survey. In contrast, those who failed the attention check had an average age of 29.4, 57% were female, and they took about a minute less (i.e., 223 seconds on average) to complete the survey. Although the amount of time spent on taking the survey is not an absolute measure of "attention", we note that the failed group still spent, on average, about 25 seconds per restaurant. We take this observation to indicate that they still gave substantial attention to their task. However, given our *a priori* decision to only use

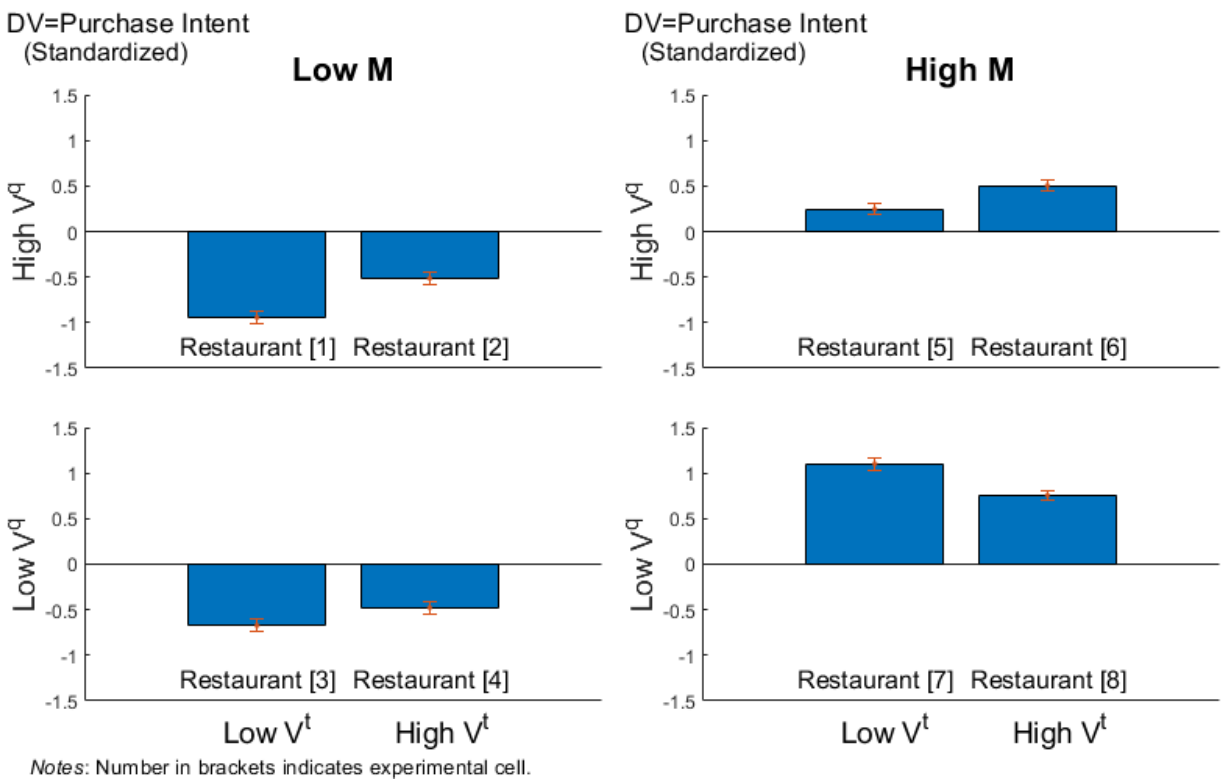

Figure 8: Purchase Intent for Restaurants in a 2 x 2 x 2 Design

"qualified" respondents, we present the results for those who passed our very strict attention test. With this noted, our hypotheses test results for the combined passed/failed groups were qualitatively similar to our reported results for the passed group only.[19] The only difference occurs when we limit our tests to the failed group, in which case we do not see all of our hypotheses supported. The additional test results from the different combinations of groups are found in in Online Appendix OD.

Before presenting our formal tests of our two hypotheses, we display the mean responses for the eight cells in Figure 8 with standard error bars. As is evident from this figure, we see increases in purchase intent for high mean cells compared to low mean cells, all else equal. Similarly, we see higher purchase intent for low V^q cells compared to high V^q cells, all else equal. However, the effect of V^t varies depending on the levels of M and V^q. With this noted, H1 involves the marginal effect of V^q across the total population and H2 states that sales increase with increases in V^t when either M is low and/or V^q is high.

[19] We conjecture that the additional noise in the failed group's responses negated the added power from almost doubling the sample size.

Table 1: Adjusted p-values under Multiple Hypothesis Testing

Alternative Hypothesis on Purchase Intent		Raw p-value	Adjusted p-value
Effect of M is positive Restaurants [1],[2],[3],[4] < [5],[6],[7],[8] (pooled)		<0.0001	<0.0001
Effect of V^q is negative Restaurants [1],[2],[5],[6] < [3],[4],[7],[8] (pooled)		<0.0001	<0.0001
Effect of V^t is more positive compared to high M & low V^q:	When M is low: Restaurants [8]−[7] < [4]−[3]	<0.0001	0.0001
	When V^q is high: Restaurants [8]−[7] < [6]−[5]	<0.0001	<0.0001
	When M is low & V^q is high: Restaurants [8]−[7] < [2]−[1]	<0.0001	<0.0001

Notes: Number in brackets indicates experimental cell.

Thus, in Table 1 we present multiple contrasts[20] where we tested the set of inequalities simultaneously (thereby avoiding the issue of alpha inflation due to multiple hypothesis tests), by using a bootstrapping method that also addresses the issue of test hypotheses being correlated due to sharing data (Westfall and Young, 1993).[21] We used 100,000 re-samples, and the adjusted p-values correspond to using the bootstrapping method. In conducting these tests, we standardized the purchase intent responses by individual across the eight restaurants to remove all individual effects related to the tendency to rate using high/low and wider/narrower responses.

Consistent with the visual view of the data, we find strong support for our hypotheses,

[20] Although not part of our stated hypotheses, we also tested if the purchase intent is higher for high mean restaurants compared to the low mean restaurants (effect of M) as a manipulation check.

[21] A bootstrapping method creates pseudo-datasets by randomly sampling with replacement from the observation data. Each randomly created pseudo-dataset represents the empirical distribution of the null of all treatments being equal. The p-values of the hypothesis tests are calculated on the re-sampled dataset, and the minimum p-value is recorded for each random pseudo-set. A large number of resampling is performed, and the adjusted p-value is calculated as the proportion of the re-sampled pseudo-p-values that are less than or equal to the raw p-value. Hence, the resampling methods implicitly accounts for all forms of correlations (inter-test and inter-variable). The resampling-class methods have been shown to have consistent Type-I errors under various correlation levels and structures (Blakesley et al., 2009), and the step-down version, which uses a subset of resamples to increase power (Holm, 1979; Shaffer, 1986; Westfall and Young, 1993), is helpful for detecting true differences while controlling the family-wise error rates across all multiple hypothesis tests under study (Romano, Shaikh, and Wolf, 2010). We refer the readers to the cited literature on multiple hypothesis testing for more details.

i.e., sales decrease with V^q across different levels of M and V^t, while the effect of V^t is more positive when M is low and/or V^q is high.[22] In addition, we note that the significance of the contrast "[8]−[7] < [6]−[5]" (which captures the positive interaction between V^q and V^t, under high M), is equivalent to testing whether the effect of V^q is less negative when V^t is high (i.e., [8]−[6] < [7]−[5]). Again, these differences are visually noticeable in Figure 8.

FIELD STUDY

Data

Classifying All Reviews

Unlike the laboratory study where we created each review, we need to determine V^q and V^t for each establishment for our field studies. This involves a multi-step process. First, we need to classify the degree to which each text review discusses vertical vs. horizontal attributes, i.e., create a measure of q_i. To do this we built two classifier algorithms for each sample. Given the voluminous quantity of text reviews for both types of establishments, we decided to use a bag-of-words model that considers words as the building blocks of textual review content[23] and a support vector machine (SVM) algorithm to classify each review in terms of its emphasis on quality and taste. We first transformed all review texts into a set of words. Common abbreviations and internet usage terms were converted to standard English, and misspelled words were corrected using a spell check program. We removed punctuation after converting the dollar sign ("$") to the word "dollars". Any meaningless "stop words" (e.g., "a", "is", "the") were removed before stemming all words into their root form (for more details, see Online Appendix OE).

Next, we used the AMT workers' classification provided for the two samples of 5,000

[22] We find these results are robust to other testing specifications, such as the step-down version of the bootstrap re-sampling method, with the heterogeneous variance instead of the homogeneity assumption (Satterthwaite approximation), and even the Bonferroni method (a more conservative test under the independence assumption).

[23] Accommodating more complex sentence structures and using deep learning algorithms might have further improved this classification. However, the texts in our study are all reviews about restaurants and hotels and consequently, there is little reason to expect the need to take into consideration the specific context of discussion when determining the classification ability of any given word. In addition, this simplifies our methodology and also makes the algorithm for training and prediction computationally efficient.

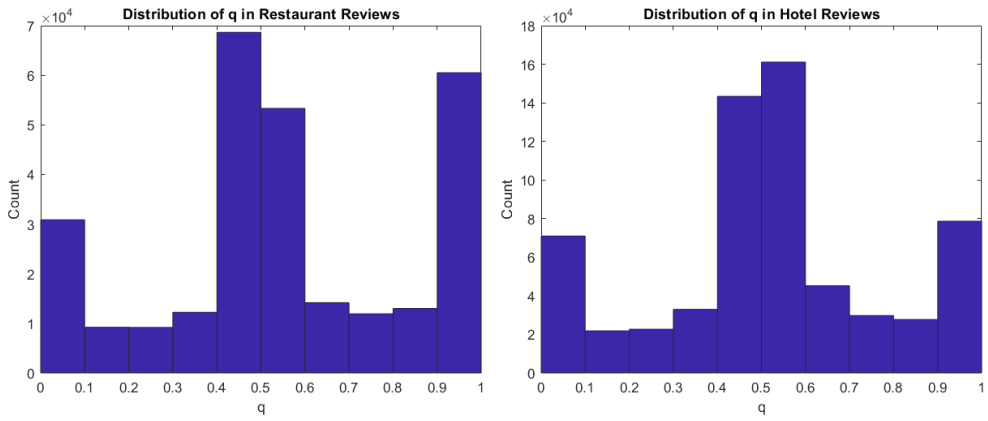

Figure 9: Distribution of q in Restaurant and Hotel Reviews

reviews to train the SVM classifiers, with a linear kernel, for each data set. Out of the entire set of words contained in each review text, we decomposed the text input answer for the phrases related to quality issues for that review into words, and then linked those words (X_q variables) to the binary response variable $Y_q = 1$, indicating the existence of quality-related content. We linked all remaining words in the review text to $Y_q = 0$, indicating no existence of quality-related content. Thus, we divided each review into sub-parts to make the training set of 10,000 observations (half of which predict $Y_q = 1$ and the other half $Y_q = 0$). We did the same for the response variable Y_t for the existence of personal taste-related topics (see Figure 4). Using these two trained SVM models, one for quality and the other for taste, we predicted the posterior probability that a review discusses quality- and taste-related content, using a logistic mapping function. Next, we scaled these two posterior probabilities to add to 1, giving us our measures of q_i and $(1 - q_i)$ for each review. We display the distribution of these predicted values in Figure 9. Due to the uncertainty in the measurement of these posterior probabilities and the observed empirical distribution of the predicted values, we binned these continuous measures, letting them take 5 discrete values: 0, .25, .5, .75, or 1, using the following bins: [0, .1), [.1, .4], (.4, .6), [.6, .9], and (.9, 1].

Our next step was to assume that the proportion of each review i devoted to quality (q_i) and taste ($1 - q_i$) is indicative of how much the quality realization and taste mismatch cost deviated from the first period consumer's prior expectations. The valence of the review content, i.e., the direction of the shock, can be determined for the most discussed content (either quality or taste) by the sign of the difference between the review rating and the

average rating of the firm. For example, if a review gives a restaurant that averages 4 stars a 3-star rating, and the review focuses more on quality, i.e., q_i>.5, then the period 2 consumer can easily infer that there was a negative deviation in quality from the consumer's prior expectation. What is less clear is whether the taste deviation (if any) was also negative, or if it was positive. The sum of the deviations must equal the total deviation in rating, but it could be the case that there was a negative shock to both quality and taste, or a very negative shock to quality that was not fully compensated by a smaller, positive shock to taste. (We note, however, that when the review is entirely about quality or taste, i.e., $q_i = 0$ or 1, there is no uncertainty—the deviation in the rating from the average rating is equal to the shock along the dimension discussed in the review.) In our main specification, we assume that consumers integrate out over this uncertainty (see Appendix for details). As a robustness check, we will assume that period 2 consumers can perfectly assess the valence of both shocks, when both quality and taste are discussed in the review.

Finally, using these quality and taste deviations, consumers (and we) can calculate for each establishment the expected variance of the ratings that results from the variation in vertical quality experiences and the variation in the horizontal taste experiences. Namely, the variance of the quality shocks (from quality expectation) would be the variance of the data generating process of quality, and likewise for taste shocks. We use these expected variances to estimate their causal effects on revenue using instrumental variables to control for endogeneity. Our instruments come from the entire review history of every reviewer who provided a relevant review for the establishments in our sample. Each Google user's landing page contains all reviews written by that user. For each business's set of review writers, we aggregate each reviewer's other ratings, which allows us to measure whether these specific reviewers are harsh or lenient. We assume that the mix of harsh and lenient reviewers over time that visit an establishment is exogenous (controlling for the baseline mix with establishment fixed effects), which allows us to create instruments for M, V^q, and V^t using these reviewers' past reviews of other establishments (the instruments being the mean of past reviews of other establishments across reviewers, and the variances across those past means).

Business Demand

We attempted to collect sales data for each restaurant and hotel in our sample. Monthly sales data for each hotel are available from the State of Texas Comptroller's office upon request. We aggregated these data to be at the quarterly level. However, revenue figures for restaurants are not normally publicly available. What is available are estimates of annual sales in a database maintained by *Infogroup*, which is considered to be one of the most comprehensive business listing databases available. This database provides detailed information about each business location including industry codes, number of employees working at the specific location, estimated annual sales, and whether it is part of a franchise. We used the *Infogroup* estimate of annual sales (in thousands of dollars) after matching their restaurant location with the restaurant found in our sample database.[24]

Summary Statistics

Table 2 displays the relevant summary statistics for our samples of restaurants and hotels, in which the unit of analysis is at the annual level for restaurants and at the quarterly level for hotels. Let Rev_{jt} indicate revenue for business j during period t. The mean of the ratings for establishment j is M_{jt} and our calculated variances are denoted with superscripts: V_{jt}^{q} is the expected variance coming from quality inconsistency and V_{jt}^{t} is the expected variance associated with taste mismatch costs. We also control for policy changes, such as renovations

[24]There are two limitations in using this database for the sales figures for restaurants. First, these sales figures are only estimates, although there is no reason to believe that the *Infogroup*'s methodology should induce a systematic bias associated with our independent variables (see Online Appendix OF for more details on the variables used to predict annual sales). Thus, using their estimated sales only adds noise to our dependent variable. Second, this database only included information for about 56% of all of the businesses in our initial sample of 7,663 restaurants which had reviews posted on the Google website (see Online Appendix OG for the exact procedure we used to match businesses across the two different databases). However, when we compared those businesses where we obtained matches and ultimately use in our demand analyses with those where we were unable to obtain sales data from the *Infogroup* database, we find the two subsamples had almost identical average ratings and average variance in the ratings. In addition, the distribution of the number of reviews was very similar. Perhaps just as importantly, when we did an analogous analysis for our hotel database and compared the annual estimates based on the *Infogroup* data with the revenue figures obtained from the State Comptroller's office, we find that the correlation of the two measures is .80 (see Figure OF.1 in Online Appendix OF). In addition, when we regressed the difference in actual and estimated hotel sales against our three parameters of interest, after controlling for firm fixed effects and time dummies (the control independent variables in our empirical model), we found these three parameters of interest are insignificantly related to the dependent variable, i.e., they are independent of the measurement error related to using the *Infogroup* data. Thus, we do not believe there is any significant issue with measurement errors or selection bias by using our estimates of restaurant sales.

Table 2: Descriptive Statistics of the Variables

	Restaurants (Yearly Panel)					Hotels (Quarterly Panel)				
Variable	Obs.	Mean	Std. Dev.	Min	Max	Obs.	Mean	Std. Dev.	Min	Max
ln sales* ($log(Rev)$)	15241	5.575	1.456	2.303	11.641	28393	12.896	0.967	4.554	17.148
Avg rating (M)	15241	4.029	0.531	1	5	28393	3.816	0.715	0.424	5
Variance due to quality (V^q)	15241	0.450	0.432	0	5.333	28393	0.565	0.478	0	8
Variance due to taste (V^t)	15241	0.311	0.306	0	4	28393	0.509	0.433	0	8
Policy change (Pol)	15241	0.022	0.146	0	1	28393	0.096	0.294	0	1
Franchise ($Fran$)	15241	0.115	0.319	0	1	28393	1	0	1	1
Within-establishment Variances										
ln sales* ($log(Rev)$)	3506	2.051	2.265	0	29.239	1629	0.073	0.267	0	5.587
Avg rating (M)	3506	0.072	0.145	0	3.684	1629	0.126	0.197	0	2.044
Variance due to quality (V^q)	3506	0.104	0.289	0	6.297	1629	0.132	0.466	0	10.049
Variance due to taste (V^t)	3506	0.054	0.136	0	2.082	1629	0.103	0.317	0	0.5

*: in thousands for restaurants only.

or changes in management (Pol_{jt}), measured by screening the review texts for such mentions, and whether the business is a franchisee ($Fran_{jt}$). In the bottom four rows, we show the within-establishment variation in the dependent variable and key regressors. Although the cross-sectional variation in the mean rating and rating variances is much larger, there is still within-firm variation that we utilize in estimation.[25]

We report the correlation among the variables in Table 3. There do not appear to be any serious signs of potential multicollinearity, other than the moderate negative association between the mean and the two variances. This negative relationship is not surprising since the average rating is approximately 4 out of 5 in both data sets and thus the distribution of ratings is skewed left. Consequently, higher variance ratings tend to have lower average mean ratings. We also note the negative correlation (across businesses) between log sales and mean rating for restaurants (but not hotels). We conjecture that this negative relationship is due, at least in part, to fast food outlets having higher sales, but lower mean ratings compared to smaller, higher end restaurants. We account for this type of relationship in our estimation by including fixed firm effects in the estimation model.

[25]To assess our estimation approach and underlying model assumptions, in addition to the concerns about using estimated restaurant revenues and the potential lack of sufficient variability in data, we run a simulation to demonstrate how data generated using similar setups can be used to recover 2SLS estimates. Please find the detailed information about this simulation analysis in Online Appendix OH.

Table 3: Correlation of Variables

	Restaurants						Hotels					
	$log(Rev)$	M	V^q	V^t	Pol	$Fran$	$log(Rev)$	M	V^q	V^t	Pol	$Fran$
$log(Rev)$	1						1					
M	-0.1143	1					0.3250	1				
V^q	-0.0548	-0.4405	1				-0.1699	-0.3564	1			
V^t	0.0468	-0.4011	0.2244	1			-0.1215	-0.3676	0.2185	1		
Pol	-0.0840	-0.0091	0.0217	0.0303	1		0.0398	0.0032	0.0229	0.0135	1	
$Fran$	0.1032	-0.3125	0.1434	0.1506	0.0195	1	.	.	.	.	.	.

The full data set includes 15,241 restaurant-year observations and 28,393 hotel-quarter observations. The distribution of the number of text reviews for both types of establishments has a long upper tail, with median 12/max 1,539 for restaurants and median 30/max 948 for hotels. Within a given establishment, we limit our attention to the 20 text reviews that appear at the top of the list on the Google reviews page at each time period (if the establishment has less than 20 reviews we look at all the available reviews). The default ordering is by what Google considers "Most Relevant", which is typically by recency but occasionally an older review appears above newer ones if it is deemed to be more helpful. We do this for two overarching reasons. First, our theoretical model is based on what consumers take away from the read reviews and consumers typically focus on the more recent reviews that appear on top when there are large numbers available. In addition, the variation in V^q and V^t over time will be limited if they capture the entire history of reviews, since any additional review will have limited impact in the calculation of these variables.

Model and Estimation

Model

We present two related econometric models. The first is analogous to our laboratory setup in that it captures the heterogeneous effect of V^t for 2x2 bins of low and high M and V^q. The second formulation directly tests the stated H2. The dependent variable in both formulations is log of revenue, for establishment j at time t, and both formulations include a number of

control variables. Our first formulation is as follows:

$$log(Rev_{jt}) = \gamma_1 M_{jt} + \gamma_2 V^q_{jt} + \gamma_3 V^t_{jt} + \gamma_4 (V^t_{jt} \times \mathbb{1}_{\text{low}M_{jt}\,\&\,\text{low}V^q_{jt}}) + \gamma_5 (V^t_{jt} \times \mathbb{1}_{\text{high}M_{jt}\,\&\,\text{high}V^q_{jt}}) + \gamma_6 (V^t_{jt} \times \mathbb{1}_{\text{low}M_{jt}\,\&\,\text{high}V^q_{jt}}) + \theta_1 Pol_{jt} + \theta_2 Fran_{jt} + \phi_j + \tau_t + \epsilon_{jt}. \quad \text{(11, 2x2 Specification)}$$

We discuss the control variables first and then detail the specific interaction formulation around the review variables. Our models allow for business-specific fixed-effects, denoted by ϕ_j, that capture each firm's intrinsic characteristics (including quality of service and taste location), which are consistent over time. Any variation over time due to economic conditions or other common macro shocks is absorbed by the time dummies, τ_t. We also include Pol_{jt} as an indicator for a management change occurring during time t, while $Fran_{jt}$ is an indicator for being a franchisee or part of a chain, and ϵ_{jt} is an idiosyncratic error. Consistent with previous literature, we take the natural logarithm of the revenue to estimate relative effects. Including the time-invariant fixed effects helps alleviate endogeneity concerns. Any remaining demand shocks that also affect our review variables and are unobserved to the econometrician are controlled for by instrumenting for the endogenous variables, namely M_{jt}, V^q_{jt}, and V^t_{jt}.

In our 2x2 specification, we divide the sample into four bins of low and high M and V^q[26] to allow for a heterogeneous effect of V^t, using three indicator variables: $\mathbb{1}_{\text{low}M_{jt}\,\&\,\text{low}V^q_{jt}}$ is the indicator variable for low M and low V^q bin, $\mathbb{1}_{\text{high}M_{jt}\,\&\,\text{high}V^q_{jt}}$ for high M and high V^q bin, and $\mathbb{1}_{\text{low}M_{jt}\,\&\,\text{high}V^q_{jt}}$ for low M and high V^q bin according to our definition. Thus, γ_3 is the effect of V^t in the high M and low V^q bin, and the effects in other bins are relative to the effect in this bin. This 2x2 specification allows us to test the effect that lowering M or increasing V^q has on the effect of V^t, from the baseline bin where the effect of V^t is the most negative. While this result is interesting, to confirm that each of M and V^q has a moderating effect on V^t, our H2 states that having low M and/*or* high V^q affects the impact that V^t has on revenue. Therefore, we also utilize an alternative specification that directly tests our

[26]Low and high M is defined using a median split. On the other hand, low and high V^q are defined relative to M because M and V^q tend to co-vary in our empirical setting of skewed rating distributions. We account for this co-variation by finding the best fit line between M and V^q, and then dividing the sample into two groups: the one with V^q higher than the predicted value relative to its M according to the best fit line, and the one with V^q lower than the predicted value.

hypotheses:

$$log(Rev_{jt}) = \gamma_1 M_{jt} + \gamma_2 V^q_{jt} + \gamma_3 V^t_{jt} + \gamma_7 (V^t_{jt} \times \mathbb{1}_{\text{low}M_{jt}\text{ or high}V^q_{jt}}) + \theta_1 Pol_{jt} + \theta_2 Fran_{jt} + \phi_j + \tau_t + \epsilon_{jt} \quad (12, \text{Direct Test of Hypotheses})$$

where $\mathbb{1}_{\text{low}M_{jt}\text{ or high}V^q_{jt}}$ is the indicator variable for being in either low M and/or high V^q condition. Similar to the 2x2 specification, γ_3 is the effect of V^t in the high M and low V^q bin where neither of the required conditions is met, and γ_7, the effect when at least one of the conditions is met, is relative to this base condition. H1 is supported if $\gamma_2 < 0$, while H2 implies $\gamma_7 > 0$.[27]

Estimation

There are a number of issues that we must address during estimation. First, our parsimonious two-period theory model assumes that second period consumers are able to observe the actual distribution of ratings across the population; in actuality, they observe a set of draws from that distribution. Thus, even if second period consumers are able to perfectly decompose the rating deviations into their quality and taste components, there will still be variation over time in both V^q and V^t based on the actual draws from the quality and taste distributions. Furthermore, we expect heterogeneity in the degree to which reviewers use more positive ratings vs. more negative ratings to express the same utility realization, reflecting how "harsh" the reviewer is when attaching a rating to a review. In our field studies, we allow for the rating left by first period consumers to have an extra additive stochastic component which reflects how harsh a specific reviewer is when assigning their ratings. Referring back to the theory model, the expected mean rating left by the first period consumers is unchanged, and the expected variance is now $V = \frac{1}{12}(r^2 + t^2 D_1^2) + V_\eta = V_q + V_t + V_\eta$, where V_η is the variance due to this extra source of variation. If second period consumers are unaware of the fact that ratings reflect this extra stochastic component, then second period consumers proceed exactly as described in the theory model. If, on the other hand, they are aware of this third source of variation, they will want to adjust their the magnitude of

[27] Note that $\gamma_3 + \gamma_7$ may still be negative since our binning may not capture the condition where M is low enough and/or V^q is high enough to generate positive sales. We discuss this more when presenting the results.

their decomposition of V into V^q and V^t. In such a case, we make the empirically supported assumption that these second period consumers will not take the effort to determine the individual reviewers' tendencies of harshness, but instead just discount their estimates of V^q and V^t when determining the underlying parameters of our model. Because this extra source of variation is expected to not change over time, our hypotheses are not affected, but the interpretation of the coefficient estimates change slightly, since the second period consumers only use the discounted estimates of the actual variance in the quality and taste distributions, while we assume in estimation that there is no such discounting. Consequently, the obtained coefficients using our approach would be underestimates of the actual effect of changes in the variances of the quality and taste distributions.

Another issue that needs to be addressed in estimation is endogeneity. There are at least two potential sources. First, even with establishment fixed effects, random shocks in unobserved quality parameters (i.e., the distribution from which quality realizations are drawn) over time could impact both ratings and sales. This would lead us to overestimate the impact of mean reviews. It also may lead to biases in the estimates of the variance of reviews, due to the observed negative correlation in the mean and the variances of reviews. Specifically, a positive correlation between any unobservable with mean reviews and sales implies a negative correlation in the unobservable and the variances of reviews, potentially leading to underestimation of the effect of review variances. A second possible source of endogeneity results from the fact that users probably do not read all of the reviews, yet our dataset does not allow us to determine which reviews they read, leading to possible measurement error. Our setting is in contrast to the one in X. Liu, D. Lee, and Srinivasan (2019), who had a dataset that allowed them to infer the reviews read. They show that incorrectly assuming all reviews are read versus using the actual reviews read leads to substantially biased estimates. To help address this concern that not all reviews are read by the second period consumers, we use (up to) the most relevant 20 reviews in estimation to calculate the variance variables, recognizing that the measured variables are still not necessarily the same as the variance variables utilized by consumers, if they read a different subset. If we assume that the variances of the reviews that are read are equal to the variances of the 20 reviews we use, plus some measurement error, then the ordinary least squares (OLS) estimates will be subject

to (potentially severe) attenuation bias.[28] This will not only lead to underestimates of the variance effects when using OLS, but also will bias the estimated effect of the mean due to the collinearity between the mean and the variances.

We address both sources of endogeneity using instrumental variables (IV) analyses. We do this by leveraging the plausibly exogenous variation in the types of consumers who previously were patrons and left reviews for the given establishment, as described above. For example, if the reviews for an establishment come from a set of consumers who generally leave lower evaluations, then we can expect the ratings for this establishment to also be lower for reasons uncorrelated with the actual quality of the reviewed establishment.[29] Note that using this reviewer information as instruments means we are implicitly assuming that 1) there is additional variation in ratings that comes from the relative "harshness" of reviewers (i.e., the degree to which ratings left by the reviewers for a business are lower or higher than the actual utility realizations), and 2) consumers of review content for a given establishment do not search through the review history of all the reviewers to assess whether they are harsh or generous reviewers. The actual instrument used for the mean rating of an establishment is $\sum m_i/N$, the average across reviewers of each reviewer's historic average mean rating, m_i, not associated with the focal establishment.

Similarly, the quality and taste variances of ratings can be instrumented by taking the variances of the deviations across reviewers from the relevant reviewers' historic mean ratings. The construction of these instruments is analogous to that of the expected variances, except that the deviations in the ratings $(s_i - \bar{s})$ are replaced with the deviations in the users' historic mean ratings $(m_i - \bar{m})$, where $\bar{m}$ is the mean as defined above. Hence, the instrument captures whether the reviewers who left reviews for an establishment are harsh or lenient raters relative to the expectation, and how much this "harshness" varies for a set of reviewers. The

[28] Huang and Sudhir (2019) also find that the downward bias due to measurement error dominates the better known upward bias due to common-method in OLS estimates, therefore leading to a significant underestimation of causal effects.

[29] Using the IV estimation also addresses possible omitted variable bias. In field data, there could be potential deviations from our theoretical model. For example, even though our analytical model assumes the location of each establishment to be fixed on the horizontal taste dimension, establishments may change their menus and thereby affect the reviews without being observed by the econometrician. Such menu turnovers, however, are unrelated to our IVs, which capture the exogenous variation due to the "harshness" user characteristics.

instruments are subject to the same integration over the unobserved shocks, as we assume that q_i is exogenous, i.e., the data generating process of quality and taste distributions are independent. In other words, which mix of quality and taste realizations are experienced by a particular consumer is independent from the endogenous demand factors of the focal establishment.

Results

Main Results

We present our estimation results[30] in Table 4 for both specifications, for both restaurant and hotel samples. We report heteroskedasticity and autocorrelation consistent (HAC) standard errors (Newey-West), assuming an AR(1) process.[31] We assume this autoregressive process to account for possible autocorrelation in the errors, as the same review may be included for calculating the variances for more than one time period. The first-stage identification test results are also included in the result tables. All individual first-stage regressions for each endogenous variable were strongly identified, and the joint weak identification tests had reasonable F statistics over ten.

We find the main effect of V^q (i.e., γ_2) is negative as hypothesized for both restaurants and hotels, although this effect is significant only for hotels at 5% (for restaurants, it is significant at 10% using the "direct test of hypotheses" specification). We find $\gamma_4, \gamma_5, \gamma_6$ and $\gamma_7 > 0$ as predicted by our second hypothesis for both restaurants and hotels, and the effects are significant at 5%. These positive coefficients imply that, compared to the baseline bin where the effect of V^t is most negative (high M and low V^q condition), the effect of V^t becomes more positive when either M is low or V^q is high. We also provide adjusted p-values that account for the worst-case bias in the test statistics due to relatively modest first-stage F statistics, as described in D. S. Lee et al. (2020). We still find support for H1 for hotels and H2 for both restaurants and hotels at 5% significance (and support for H1 for restaurants at 10%) using the adjusted p-values in the specification that directly tests for both hypotheses.

[30] First-stage estimates are included in Online Appendix OI.

[31] We also ran analyses with longer serial correlation, and found the results did not materially change.

Table 4: Main Estimation Results

			Restaurants (Yearly Panel)								Hotels (Quarterly Panel)							
			2x2 Specification				Direct Test of Hypotheses				2x2 Specification				Direct Test of Hypotheses			
			Newey			*Adjusted*	Newey			*Adjusted*	Newey			*Adjusted*	Newey			*Adjusted*
$log(Rev)$		Prediction:	Coef.	Std. Err.	p-value	p-value	Coef.	Std. Err.	p-value	p-value	Coef.	Std. Err.	p-value	p-value	Coef.	Std. Err.	p-value	p-value
M	γ_1		-0.1297	(0.1718)	0.225	0.225	-0.1025	(0.1414)	0.234	0.234	-0.0238	(0.0370)	0.260	0.260	-0.0160	(0.0306)	0.300	0.300
V^q	γ_2	$\gamma_2 < 0$	-0.4216	(0.3459)	0.111	0.111	-0.3342	(0.2502)	0.091*	0.091*	-0.1641	(0.0764)	0.016**	0.039**	-0.1359	(0.0558)	0.007***	0.021**
V^t	γ_3		-1.2938	(0.5710)	0.012**	0.042**	-1.1938	(0.4880)	0.007***	0.024**	-0.1696	(0.0922)	0.033**	0.050*	-0.1673	(0.0894)	0.031**	0.038**
$V^t \times \mathbb{1}_{\text{low}M \,\&\, \text{low}V^q}$	γ_4	$\gamma_4 > 0$	0.6908	(0.3362)	0.020**	0.048**					0.1031	(0.0509)	0.021**	0.043**				
$V^t \times \mathbb{1}_{\text{high}M \,\&\, \text{high}V^q}$	γ_5	$\gamma_5 > 0$	0.7257	(0.3959)	0.033**	0.057*					0.1378	(0.0768)	0.036**	0.052*				
$V^t \times \mathbb{1}_{\text{low}M \,\&\, \text{high}V^q}$	γ_6	$\gamma_6 > 0$	0.9880	(0.5597)	0.039**	0.060*					0.1913	(0.0939)	0.021**	0.043**				
$V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$	γ_7	$\gamma_7 > 0$					0.7049	(0.3281)	0.016**	0.032**					0.1320	(0.0609)	0.015**	0.028**
Pol			-0.1230	(0.0636)	0.027**	0.053*	-0.1318	(0.0608)	0.015**	0.031**	0.0008	(0.0065)	0.448	0.448	0.0013	(0.0064)	0.420	0.420
$Fran$			0.0320	(0.0801)	0.345	0.345	0.0231	(0.0792)	0.385	0.385								
Firm Fixed Effects			Included				Included				Included				Included			
Time Fixed Effects			Included				Included				Included				Included			
Num Obs.			15241				15241				28393				28393			
Num Panelists			3506				3506				1629				1629			
Avg Obs./Panelist			4.3				4.3				17.4				17.4			
First-stage Underidentification Test:			χ^2=25.05		p-val=0.0000		χ^2=41.62		p-val=0.0000		χ^2=24.85		p-val=0.0000		χ^2=33.95		p-val=0.0000	
First-stage Weak Identification Test:			F=12.51				F=27.83				F=17.63				F=36.70			

***: significant at 1%, **: significant at 5%, and *: significant at 10% (**one-tailed tests**).

Placebo Test

Future information should not affect past sales. We conduct this placebo test by regressing sales on future review statistics. For each unit of analysis in our main estimation, we calculate M, V^q, and V^t using future reviews, rather than past posted reviews. In order to match the amount of information available for each establishment, we use (up to) the next n reviews posted in the future to calculate the mean and the variances, where n is the number of reviews used in the main estimation. We show that information from future reviews do not predict past sales, and the estimation results are presented in Table 5. The placebo test results tell us that reverse causality is not at play here, since otherwise past sales would be related to future reviews.[32]

Robustness Checks

As mentioned earlier, it could also be the case that second period consumers can perfectly infer the valence of both the quality and taste deviations from the review content for all reviews. In this case, they do not need to calculate the expected V^q and V^t, as they can calculate them with certainty, implying a difference between the consumer's information and that of the researcher. In this situation we must integrate over our uncertainty in the estimation procedure. We do this by using the same set of 1000 simulations of the realizations of the quality and taste shocks described in Appendix . However, instead of calculating the consumers' expected V^q and V^t by averaging across these simulations, we estimate the empirical model for every simulated set of draws, yielding 1000 independent estimates of our coefficients, each based on a different imputed value of V^q and V^t (using HAC errors as before). Following the procedure recommended by Marshall et al. (2009), we calculate the mean of each of these coefficients and the associated variances of these 1000 estimated coefficients. Finally, we include the imputation uncertainty within our standard errors using Rubin's rule (Rubin (1987)) to determine the overall variance that captures model fit as well as missing data uncertainty. While the standard errors are larger under

[32]Our identification strategy also prevents reverse causality from driving the effects in our main results, since the instruments are unrelated to our dependent variables (i.e., high sales do not cause the reviewers to become more harsh or lenient).

Table 5: Placebo Test Results

			Restaurants (Yearly Panel)								Hotels (Quarterly Panel)							
			2x2 Specification				Direct Test of Hypotheses				2x2 Specification				Direct Test of Hypotheses			
			Newey			*Adjusted*	Newey			*Adjusted*	Newey			*Adjusted*	Newey			*Adjusted*
$log(Rev)$		Prediction:	Coef.	Std. Err.	p-value	p-value	Coef.	Std. Err.	p-value	p-value	Coef.	Std. Err.	p-value	p-value	Coef.	Std. Err.	p-value	p-value
M	γ_1		0.0179	(0.0367)	0.312	0.312	0.0115	(0.0339)	0.367	0.367	0.0261	(0.0560)	0.320	0.320	0.0189	(0.0453)	0.338	0.338
V^q	γ_2	$\gamma_2 < 0$	0.0712	(0.0648)	0.136	0.136	0.0428	(0.0480)	0.186	0.186	0.0934	(0.1169)	0.212	0.212	0.0750	(0.0851)	0.189	0.189
V^t	γ_3		0.1920	(0.1694)	0.129	0.129	0.1948	(0.1563)	0.106	0.106	0.0828	(0.1407)	0.278	0.278	0.0941	(0.1437)	0.256	0.256
$V^t \times \mathbb{1}_{\text{low}M\,\&\,\text{low}V^q}$	γ_4	$\gamma_4 > 0$	-0.1007	(0.1391)	0.235	0.235					-0.0430	(0.0760)	0.286	0.286				
$V^t \times \mathbb{1}_{\text{high}M\,\&\,\text{high}V^q}$	γ_5	$\gamma_5 > 0$	-0.2061	(0.1210)	0.044**	0.056*					-0.1111	(0.1389)	0.212	0.212				
$V^t \times \mathbb{1}_{\text{low}M\,\&\,\text{high}V^q}$	γ_6	$\gamma_6 > 0$	-0.2041	(0.1802)	0.129	0.129					-0.0886	(0.1403)	0.264	0.264				
$V^t \times \mathbb{1}_{\text{low}M\,\text{or}\,\text{high}V^q}$	γ_7	$\gamma_7 > 0$					-0.1317	(0.1284)	0.153	0.153					-0.0699	(0.0964)	0.234	0.234
Pol			0.0246	(0.0241)	0.153	0.153	0.0220	(0.0242)	0.182	0.182	0.0014	(0.0067)	0.418	0.418	0.0013	(0.0067)	0.425	0.425
$Fran$			0.0214	(0.0744)	0.387	0.387	0.0284	(0.0734)	0.350	0.350								
Firm Fixed Effects			Included				Included				Included				Included			
Time Fixed Effects			Included				Included				Included				Included			
Num Obs.			13410				13410				26990				26990			
Num Panelists			3051				3051				1581				1581			
Avg Obs./Panelist			4.4				4.4				17.1				17.1			
First-stage Underidentification Test:			χ^2=28.77 p-val=0.0000				χ^2=32.24 p-val=0.0000				χ^2=5.48 p-val=0.0192				χ^2=8.75 p-val=0.0031			
First-stage Weak Identification Test:			F=18.13				F=33.68				F=4.14				F=8.27			

***: significant at 1%, **: significant at 5%, and *: significant at 10% (**one-tailed tests**).

this alternative data generating assumption due to the added researcher uncertainty, we find that qualitative interpretations of the results do not materially change, and the hypothesized effect of the variances continue to hold with significance (see Table OJ.1 in Online Appendix OJ).

We also report in Table OJ.2 in Online Appendix OJ the estimation results assuming all quality and taste deviations to be of the more general "same sign" type, as an additional robustness check. Since the opposite signed deviations represent less likely scenarios with larger magnitudes of deviations, we verify that the significant effects of the variances are not driven by the extreme values simulated based on our distributional assumptions on the deviations. Estimation results are robust to plausible variations in distributional parameters for quality and taste deviations, and even when we assume them to be of the same sign for all reviews (where no distributional assumption is needed), the significant effects of the variances continue to hold with the same significance level.

DISCUSSION AND CONCLUSION

This paper extends the existing literature on online reviews by considering the effect of the rating variance from two sources, within-firm stochastic quality and across-consumer taste mismatch costs, and in doing so broadens the applicability of such inquiries to a large number of service industries where service quality varies across consumer experiences. We find the results of a controlled laboratory experiment and two field studies are compatible with the theoretical predictions, based on our assumed process of how reviews left by past customers influence the purchase decision of future customers. The overarching takeaway from both the theoretical model and empirical findings is that the two sources of variance interact and thus, their marginal effects depend on the level of the other. The effect on future demand for taste variance can either be positive or negative, depending on the level of quality variance, while the effect of quality variance on future demand is always negative, but marginally smaller with high levels of taste variance.

Our analytic model explicitly delineates how customers determine what to include in their text reviews as well as how prospective customers interpret the information in order to gauge

a) the average quality level and the degree of uncertainty in quality, and b) the importance of mismatch costs associated with the product offering. Although we find strong support for the predictions coming from this analytic model, we note that we only directly test for one element associated with the process, namely that consumers can reliably differentiate between statements concerning quality and taste. Thus, like the analogy of a pool player knowing how to bounce a shot off the side of the pool table to sink a ball into the pouch without using exact geometry, we do not assume customers strictly follow our assumed process. However, we believe the predictions coming from our model provide important prescriptive validity. With this noted, we next discuss some implications that flow from our model and results.

Our two-period analytical model is silent on the long term implications of the effect of the reviews. In fact, it implicitly assumes that all of the underlying parameters are known by the start of the second period and thus V^q and V^t are known and fixed. However, empirically we find these two review variables vary over time. This variation could come from multiple sources: the first is the stochastic nature of the quality realizations the first period consumers receive, since second period consumers do not observe the distribution from which the ratings are drawn, but only a set of draws from that distribution. The second is the relative harshness of the first period reviewers. A third source of variation could come from firms shifting their quality parameters (e.g., investing in quality) or taste positioning. The first two sources of variation are exogenous, and we leverage the second in our instrumental variables estimation strategy. The third source of variation is what leads to long-term implications of the effect of these two variances. We focus on the most relevant recent reviews to calculate the review parameters instead of the complete history of all the reviews, since consumers rely more on recent reviews (because they feel they more accurately reflect the current characteristics of the focal establishment). This leads to varying incentives for the firm to change its quality and taste variances, depending on the historical set of reviews. Lower mean reviews can potentially lead firms to market to a broader set of consumers, leading to greater taste variance and potential increases in sales coming from this larger market. Similarly, hotels that market only to a niche audience, and have lower taste variance, may have more incentive to reduce the variation in quality realizations by concentrating on reducing the lower end of

the service quality distribution, since the marginal effect of V^q is greater with low V^t. In addition, since mean ratings are reported by review sites for the entire set of reviews, ongoing changes in quality or positioning will be reflected more, initially, in the review content itself, implying that these variances take on added importance to both the firm and the consumer.

Firms also can learn from monitoring the changes in V^q and V^t over time, since the most recent V^t and M determine the firm's average quality of service, as perceived by the customers, while the most recent V^q determines the variability of the current service quality. Thus, although it is well established that variance in service quality can be harmful for repeat business, our theoretical and empirical results show that higher quality variance positively moderates the effect of taste variance and because these two variances interact, the converse is also true. Thus, increases in V^t can result in attracting new customers because the average quality level can be inferred to be higher (i.e., the positive effect of V^t on demand) when quality variance is large. Similarly, as shown empirically in our results, low values of V^t increase the marginal effect of increasing the reliability of the establishment's service quality. Thus, the net impact of both types of variation is not straight-forward but is tempered by the level of the other variation.

Collectively, these findings suggest that firms should track the source of the total variance in their ratings, as the two types of variance have different implications for future demand. Using textual review information, an establishment can monitor the source of variance in reviews and if it is due to customers' heterogeneous tastes, the establishment can further investigate if such variance is helpful or whether it should work to realign its offerings to better fit the taste of the target customers. On the other hand, if there is large quality variance in customer feedback, the firm should determine if improving service consistency by reducing negative experiences will be cost effective, noting that such a change may nullify the positive lift to sales from taste discovery. Our findings should generalize to other familiar product categories consisting of feature-based and experiential attributes, which can be decomposed into vertical and horizontal dimensions. In addition, our methodology presents a simple and practical implementation strategy to scan, learn, and track the information contained in online text reviews.

Our findings also have implications for the review platforms, which may enhance user

experience by providing the summaries of the quality- and taste-related review content. Some platforms nowadays provide key information, such as positively and negatively valenced keywords, but since consumers utilize quality and taste information in a different way in making purchase decisions, it would be helpful to present each type of content separately (or perhaps positive/negative keywords for each dimension).

While our study's data collection for Google reviews provides exhaustive coverage of available review information on the largest search website, it also has limitations since consumers also use other online sites to view customer reviews, and in some industries, there have been other popular sites, such as Yelp for restaurants and TripAdvisor for hotels. This limitation was circumvented with our laboratory study, since the provided reviews were the only source of available information and independent of the specific platform.[33] Regarding generalizability, the results may only apply to familiar category of products, as consumers may not be capable of differentiating vertical and horizontal attributes if the product class is not well-understood by most consumers. Additionally, although our results hold for our samples in general, we believe they are most applicable to early reviews in new markets, since it is these markets where consumers have the most diffuse prior beliefs and are most likely to seek additional information.

An interesting extension would be to use our process model to investigate the effect of reviews when conditions change and new consumers use the reviews to set their priors before getting a service experience. Do the consumers' prior beliefs affect the way they interpret the text information, i.e., confirmatory bias alters the updating process? Does this altered process affect subsequent reviews and thus subsequent sales? Such an exploration can provide insights on the impact of early and more recent reviews, as well as the characteristics of the reviewers and inferences about them. This type of exploration is similar to the work of Bondi (2019) who investigates a problem of self-selection. Effect of reviews on demand under such biases would be an interesting empirical question. We leave such extensions on the incomplete information scenarios to future studies. A normative study on how firms should respond to reviews to enhance revenues would also be an interesting extension.

[33]Even in our field studies, although we were unable to control for the external information on other platforms, our results using the instruments still stand the test of significance, under the assumption that our instruments are independent from external information.

References

Anderson, Michael and Jeremy Magruder (2012). "Learning from the crowd: Regression discontinuity estimates of the effects of an online review database". In: *The Economic Journal* 122.563, pp. 957–989.

Blakesley, Richard E et al. (2009). "Comparisons of methods for multiple hypothesis testing in neuropsychological research." In: *Neuropsychology* 23.2, pp. 255–264.

Bondi, Tommaso (2019). "Alone, together: Product discovery through consumer ratings". In: *Available at SSRN 3468433*.

Boulding, William et al. (1993). "A dynamic process model of service quality: from expectations to behavioral intentions". In: *Journal of Marketing Research* 30.1, pp. 7–27.

Chaudhuri, Surajit et al. (2003). "Robust and efficient fuzzy match for online data cleaning". In: *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*. ACM, pp. 313–324.

Chen, Pei-Yu, Shin-yi Wu, and Jungsun Yoon (2004). "The impact of online recommendations and consumer feedback on sales". In: *Proceedings of the International Conference on Information Systems*, pp. 711–724.

Chen, Peiyu et al. (2021). "Measuring product type and purchase uncertainty with online product ratings: A theoretical model and empirical application". In: *Information Systems Research* 32.4, pp. 1470–1489.

Chevalier, Judith A and Dina Mayzlin (2006). "The effect of word of mouth on sales: Online book reviews". In: *Journal of Marketing Research* 43.3, pp. 345–354.

Clemons, Eric K, Guodong Gordon Gao, and Lorin M Hitt (2006). "When online reviews meet hyperdifferentiation: A study of the craft beer industry". In: *Journal of Management Information Systems* 23.2, pp. 149–171.

Duan, Wenjing, Bin Gu, and Andrew B Whinston (2008a). "Do online reviews matter? An empirical investigation of panel data". In: *Decision Support Systems* 45.4, pp. 1007–1016.

— (2008b). "The dynamics of online word-of-mouth and product sales—An empirical investigation of the movie industry". In: *Journal of Retailing* 84.2, pp. 233–242.

Harbaugh, Rick, John Maxwell, and Kelly Shue (2016). "Consistent good news and inconsistent bad news". In: *Wokring Paper*.

Holm, Sture (1979). "A simple sequentially rejective multiple test procedure". In: *Scandinavian Journal of Statistics*, pp. 65–70.

Hu, Nan, Paul A Pavlou, and Jie Jennifer Zhang (2017). "On self-selection biases in online product reviews". In: *MIS Quarterly* 41.2, pp. 449–471.

Huang, Guofang and K Sudhir (2019). "The causal effect of service satisfaction on customer loyalty". In: *Available at SSRN: https://dx.doi.org/10.2139/ssrn.3391242*.

Lee, David S et al. (2020). "Valid t-ratio Inference for IV". In: *arXiv preprint arXiv:2010.05058*.

Liu, Xiao, Dokyun Lee, and Kannan Srinivasan (2019). "Large-scale cross-category analysis of consumer review content on sales conversion leveraging deep learning". In: *Journal of Marketing Research* 56.6, pp. 918–943.

Liu, Yong (2006). "Word of mouth for movies: Its dynamics and impact on box office revenue". In: *Journal of Marketing* 70.3, pp. 74–89.

Luca, Michael (2016). "Reviews, reputation, and revenue: The case of Yelp. com". In: *Harvard Business School NOM Unit Working Paper No. 12-016.* Available at SSRN: http://dx.doi.org/10.2139/ssrn.1928601.

Marshall, A et al. (2009). "Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines". In: *BMC Medical Research Methodology* 9.57.

Meyer, Robert J (1981). "A model of multiattribute judgments under attribute uncertainty and informational constraint". In: *Journal of Marketing Research* 18.4, pp. 428–441.

Navarro, Gonzalo (2001). "A guided tour to approximate string matching". In: *ACM computing surveys (CSUR)* 33.1, pp. 31–88.

Navarro, Gonzalo et al. (2001). "Indexing methods for approximate string matching". In: *IEEE Data Eng. Bull.* 24.4, pp. 19–27.

Romano, Joseph P, Azeem M Shaikh, and Michael Wolf (2010). "Multiple testing". In: *The New Palgrave Dictionary of Economics*, pp. 1–4.

Rozenkrants, Bella, S Christian Wheeler, and Baba Shiv (2017). "Self-expression cues in product rating distributions: When people prefer polarizing products". In: *Journal of Consumer Research*.

Rubin, Donald B (1987). *Multiple imputation for survey nonresponse*. New York: Wiley.

Rust, Roland T et al. (1999). "What you don't know about customer-perceived quality: The role of customer expectation distributions". In: *Marketing Science* 18.1, pp. 77–92.

Shaffer, Juliet Popper (1986). "Modified sequentially rejective multiple test procedures". In: *Journal of the American Statistical Association* 81.395, pp. 826–831.

Sriram, S, Pradeep K Chintagunta, and Puneet Manchanda (2015). "Service quality variability and termination behavior". In: *Management Science* 61.11, pp. 2739–2759.

Sun, Monic (2012). "How does the variance of product ratings matter?" In: *Management Science* 58.4, pp. 696–707.

Tucker, Catherine and Juanjuan Zhang (2011). "How does popularity information affect choices? A field experiment". In: *Management Science* 57.5, pp. 828–842.

Vermeulen, Ivar E and Daphne Seegers (2009). "Tried and tested: The impact of online hotel reviews on consumer consideration". In: *Tourism Management* 30.1, pp. 123–127.

West, Patricia M and Susan M Broniarczyk (1998). "Integrating multiple opinions: The role of aspiration level on consumer response to critic consensus". In: *Journal of Consumer Research* 25.1, pp. 38–51.

Westfall, Peter H and S Stanley Young (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Vol. 279. John Wiley & Sons.

Zhang, Xiaoquan and Chrysanthos Dellarocas (2006). "The lord of the ratings: is a movie's fate is influenced by reviews?" In: *ICIS 2006 Proceedings*. Paper 117.

Zhu, Feng and Xiaoquan Zhang (2010). "Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics". In: *Journal of Marketing* 74.2, pp. 133–148.

Zimmermann, Steffen et al. (2018). "Decomposing the variance of consumer ratings and the impact on price and demand". In: *Information Systems Research* 29.4, pp. 984–1002.

Appendix

Quality and Taste Deviations

Once each review is classified in terms of the amount of content devoted to quality- and taste-related content, we calculate the two partial variances of the ratings for each firm as described here. Let s_i denote the rating associated with the ith review and $\bar{s}$ is the total mean of all text review ratings for that establishment at that time, and define Δ_i as the observed deviation of s_i from $\bar{s}$. The rating given by the reviewer is $s_i = v_i - tx_i$. The quality realization can be written as a deviation from the expectation, $v_i = E[v|f(\bar{v}, r)] + \Delta_i^q$, and the taste realization can similarly be written as: $-tx_i = E[-tx_i|g(t)] + \Delta_i^t$, in which $\Delta_i^q + \Delta_i^t = \Delta_i$. Once Δ_i^q and Δ_i^t are known for all text reviews, the quality and taste variance can be calculated as the variance of the set of Δ_i^q's and Δ_i^t's, respectively. We next describe how the Δ_i^q's and Δ_i^t's are determined.

Recall that we assume that the content of the review (i.e., the amount of the review focused on quality vs. taste), represented by q_i, reflects the relative deviations from the consumer's expectations. If q_i is equal to 0 or 1 for review i, then the total deviation, Δ_i, is attributed to taste or quality, respectively, and thus can be associated with either V^t or V^q. If $q_i = \frac{1}{2}$, i.e., the review discusses quality and taste in equal proportions, then $\Delta_i^q = \Delta_i^t = \frac{1}{2}\Delta_i$ when $\Delta_i \neq 0$. If $\Delta_i = 0$, i.e., the rating is equal to the mean rating, then either $\Delta_i^q = \Delta_i^t = 0$ or $\Delta_i^q = -\Delta_i^t$, i.e., there are an infinite number of possible quality and taste shocks that can explain both the observed rating and q_i, but they need to be of equal magnitude. If $0 < q_i < 1$ and $q_i \neq \frac{1}{2}$, then there are two possible sets of deviations:

$$\Delta_i^q = q_i\Delta_i \quad and \quad \Delta_i^t = (1 - q_i)\Delta_i; \quad \text{OR} \tag{A.1}$$

$$\Delta_i^q = \frac{-q_i}{1 - 2q_i}\Delta_i \quad and \quad \Delta_i^t = \frac{1 - q_i}{1 - 2q_i}\Delta_i. \tag{A.2}$$

The first solution is when the quality and taste deviations are in the same direction, and the second occurs when they are in opposite directions, in which case the deviation of the more talked about factor (quality or taste) has the same sign as the overall deviation, and the sign of the less talked about deviation has the opposite sign.

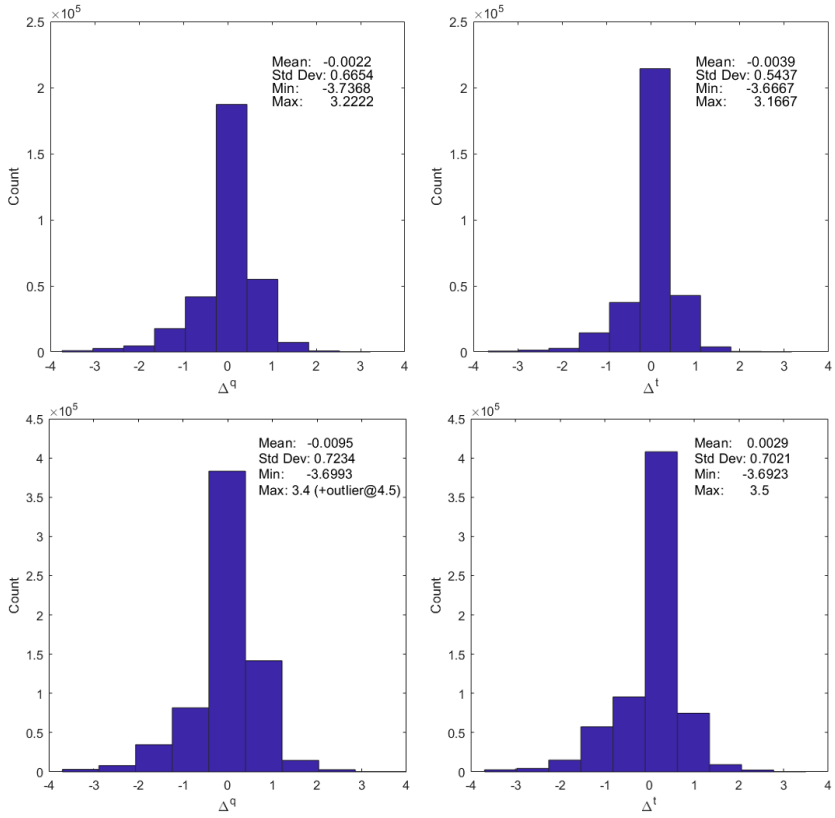

Figure A.1: Distribution of Quality and Taste Deviations in Restaurant (top) and Hotel (bottom) Reviews

In our main specification, we assume that consumers cannot distinguish between the two solutions (or in the case of $q_i = \frac{1}{2}$ and $\Delta_i = 0$, the infinite solutions). However, they may infer the sign of the more talked about deviation when $q_i \neq \frac{1}{2}$ (as the valence of the main discussion should be obvious), but this is not sufficient to assess whether the less talked about deviation is of the same or opposite valence. Hence, we assume that consumers form expectations over the unobserved deviations. To integrate over the unobserved shocks, we (and consumers) draw the likelihood of deviations from a normal distribution. In other words, small deviations from the expectation are highly likely to occur (for both quality and taste), while large deviations are less likely. We infer the distributional parameters from the observed Δ_i^q's and Δ_i^t's from the set of reviews with known Δ_i^q and Δ_i^t (i.e., those reviews with q_i equal to 0 or 1, or $q_i = \frac{1}{2}$ and $\Delta_i \neq 0$). As shown in Figure A.1, the distribution of both Δ^q and Δ^t is approximately normal with mean 0 and variance of .6 for the restaurant reviews and 0 and .7 for the hotel reviews.

Online Appendices

OA Model: Illustration

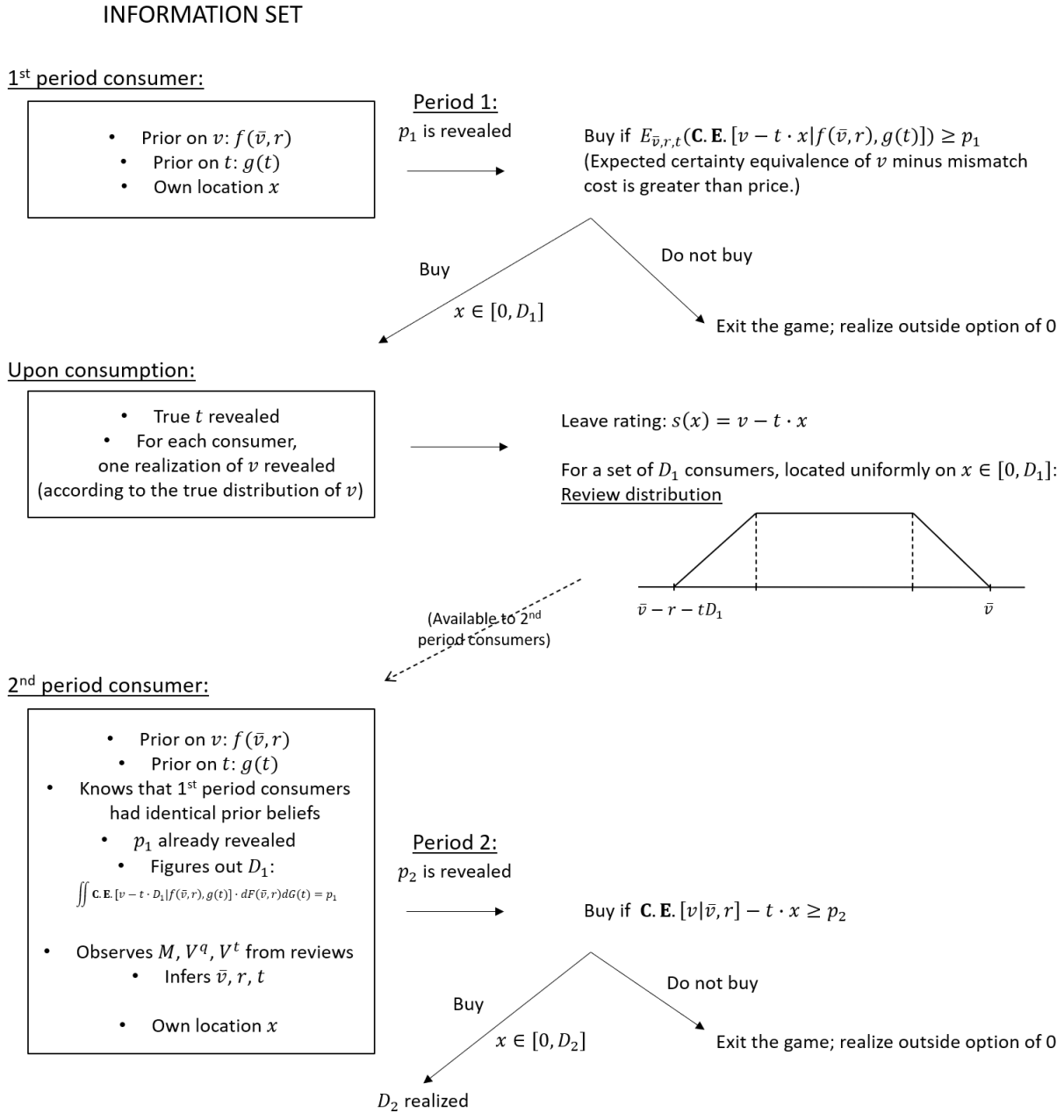

Figure OA.1: Information Flow in the Consumer Decision Process

OB Amazon Mechanical Turk Surveys

We present the information related to the classification surveys for the restaurant reviews; the procedure was analogous for hotel review classification surveys, except for the noted information at the end.

Survey Instrument

Respondents were first instructed to read four paragraphs explaining the task and the broad definition of quality-related issues and personal taste-related issues that may be present in online consumer reviews (see Figure OB.1).

They then sequentially saw 11 reviews (ten actual reviews + one attention check review) along with a set of questions about that review that appeared in an identical format. An example question is shown in Figure OB.2; this specific example contains the dummy review with the attention check phrases, instructing survey respondents to select specific answers. Note that this instruction was embedded within the review and thus required the worker to pay careful attention to the text or otherwise be disqualified. The normal questions are filled with actual reviews to measure. The button-type Yes/No or multiple-choice questions are required for submission, while the text input answers are optional.

Figure OB.1: Survey Instruction

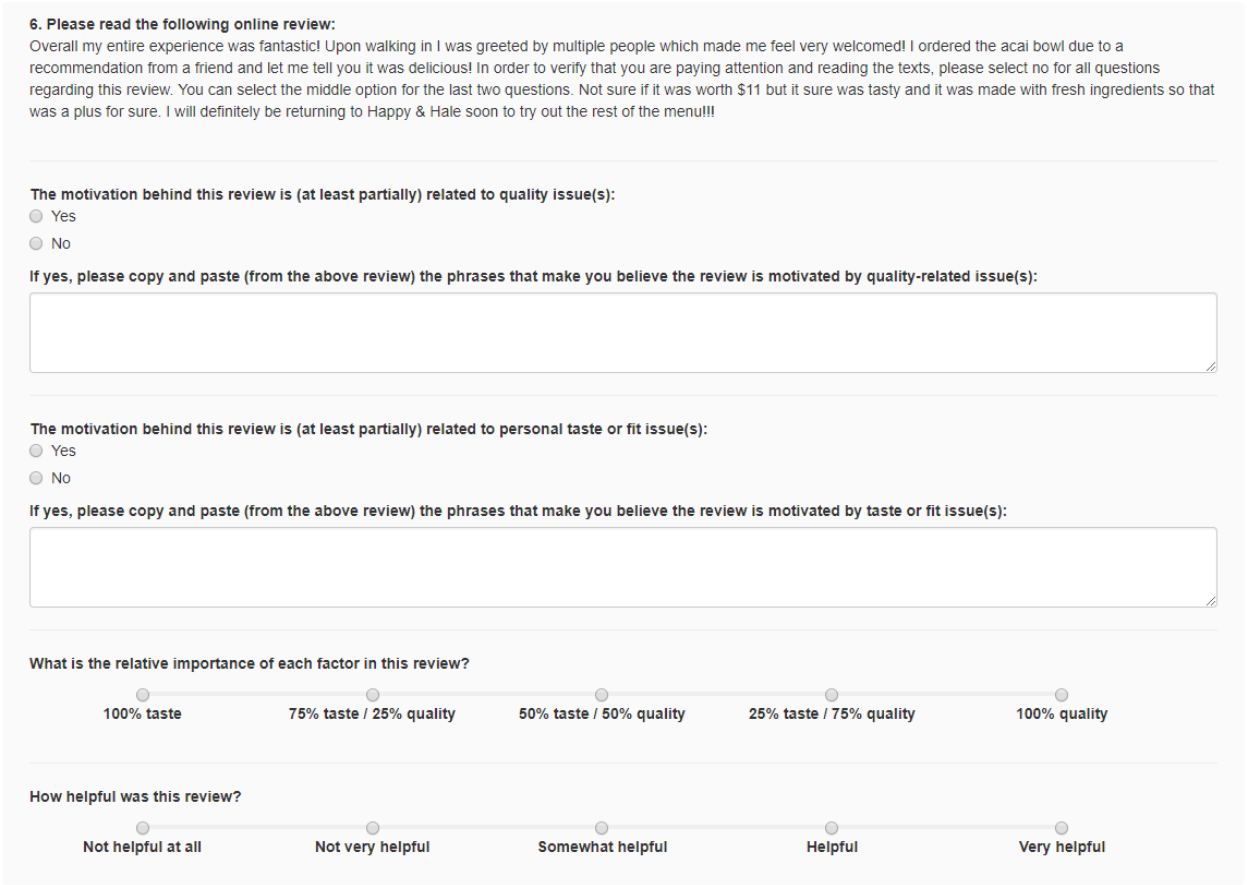

Figure OB.2: Survey Question

Survey Administration Strategies

In order to obtain high-quality responses, we employed the following strategies:

- We restricted the target AMT workers to have completed more than 100 tasks with over 97% approval rate. We did not restrict the workers to be from the U.S. This is because the user base for San Francisco restaurants may include a significant portion of foreign travelers as well. We also actively screened for the consistency in survey answers to make sure that the respondents understand English and are answering the questions asked.

- During a set of small-size pretests of the survey, we observed the following:

  – There was some confusion about selecting only negative phrases; hence, a sentence was added to the instructions so it was clear either positive or negative comments can be related to both quality or personal taste context.

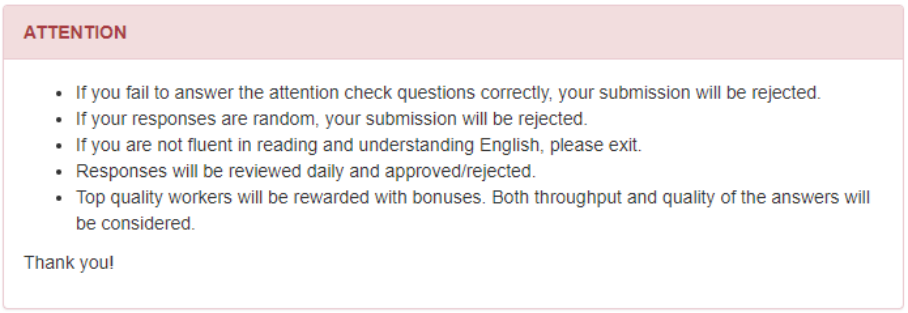

Figure OB.3: Survey Attention

  - Survey duration was also monitored during the pretest to adjust the payment level in order to be attractive to top quality workers. For first-time takers, the survey took around 15-20 minutes. For repeat takers, they took less time (6-15 min) since they were already familiar with the instruction. Any submitted work under five minutes was taken to be evidence of fraudulent activity.
  - After discovering a high rate of fraudulent attempts to submit random answers during pretest(by bots and humans), an attention box was placed at the top of the survey even before the consent form (Figure OB.3). Such dishonest attempts went down (from over 30% for pretests to under 10% for the actual surveys), after we declared clear intent to filter out those dishonest submissions and actually carried out to reject and/or block those workers.

- In addition, each survey contained one attention check question embedded somewhere in-between ten actual questions. See Figure OB.2 for the exact wording of this attention check. The placement of the attention check question varied from 6th, 7th, to 8th question. We pulled five random text reviews for some restaurants not in our sample, and used them to encase our attention check phrase. All submissions that failed to answer the attention check questions correctly were rejected, and the survey was re-assigned to another worker.

- Any submissions that show clearly meaningless responses were also rejected, and these workers were blocked to prevent them from taking our other surveys. Examples of such dishonest attempts are: 1) The worker-bot selects all Yes's and the first choice for the multiple choice questions and leaves all text input empty; 2) The worker selects random

answers for Yes/No and multiple choice questions so that their Yes/No answers do not match closely to the relativeness multiple choice answer; 3) The worker selects the first sentence of the review and copies into the text input for quality-related content, and copies the last sentence of the review into the text input for personal taste-related content, for all ten questions. Any other recognizable pattern was also rejected and these surveys were re-assigned to other workers.

- For the reviews that are rated multiple times (for demonstration of the inter-rater reliability), we used the median number of mentions for each word contained in quality- and personal taste-related text inputs.

- Overall, 197 distinct workers provided responses to 609 surveys (each containing ten reviews).

Notes for Hotel AMT Surveys

- The example paragraph in the instruction read:

  - For example, a review that states "This hotel has some serious sanitary issues; there was a mold growing on the carpet in my room." is about the quality of the hotel, while a review that states "This hotel has some great kid-friendly amenities and features, which were perfect for my family with two little kids." is motivated by personal preference. Statements could be positive or negative, for either contexts.

- We restricted the workers to be from the U.S., since Texan hotels are usually not tourist hotspots for international travelers.

- The placement of the attention check question varied from 5th to 9th question.

- Overall, 218 distinct workers provided responses to 555 surveys (each containing ten reviews).

OC Laboratory Experiment

Manipulated Reviews – Full

Figures OC.1 and OC.2 show the reviews that were manipulated in the experiment. All restaurants are shown to serve similar menu items and be in the same price category. Service quality and wait times were used to vary the quality variances, and spiciness of the food was used to vary the taste variance. All reviews were written to be within a similar range of length.

Survey Instrument

Pretest Survey. Figure OC.3 presents the instructions given at the beginning of the pretest survey. Figure OC.4 shows the actual questions used in the pretest survey. Respondents first were given instructions about the definition of quality and taste in our context. Then the AMT respondents were asked two questions about the amount of quality variance and taste variance after reading the five reviews associated with each of the eight restaurants. The order of the two questions, the eight restaurants, and the five reviews for each restaurant were randomized among the respondents. As each respondent was required to read 40 different reviews, a task requiring considerable cognitive effort, we embedded within these eight restaurants a screening question to determine if the respondents were paying enough attention to the task. Out of 75 respondents, 45 passed the screening task and thus were included in the manipulation check analyses.

The cell means for the pretest survey are shown in Table OC.1 and Table OC.2, for each question measuring the perception of quality variance and the perception of taste variance, respectively. The survey answers were standardized by individual across the eight restaurants, so each respondent's answers have a mean of 0 and a standard deviation of 1.

Main Survey. Figure OC.5 presents the instructions given at the beginning of the main experiment survey. Figure OC.6 shows the actual questions used in the main survey. In addition to the main purchase intent question, we also asked two subsequent questions that measure the two components that contribute to the purchase decision—the consistency of the

Table OC.1: Cell Means for Pretest Survey: Perception of Quality Variance

	Low M		High M	
High V^q	Restaurant [1] $V^q = 0.4845$	Restaurant [2] $V^q = 0.7481$	Restaurant [5] $V^q = 0.4716$	Restaurant [6] $V^q = 0.2688$
Low V^q	Restaurant [3] $V^q = -0.4441$	Restaurant [4] $V^q = 0.1412$	Restaurant [7] $V^q = -1.1695$	Restaurant [8] $V^q = -0.5341$
	Low V^t	High V^t	Low V^t	High V^t

Table OC.2: Cell Means for Pretest Survey: Perception of Taste Variance

	Low M		High M	
High V^q	Restaurant [1] $V^t = -0.2819$	Restaurant [2] $V^t = 0.8432$	Restaurant [5] $V^t = -0.2539$	Restaurant [6] $V^t = 0.4584$
Low V^q	Restaurant [3] $V^t = -0.5123$	Restaurant [4] $V^t = 0.6726$	Restaurant [7] $V^t = -1.0874$	Restaurant [8] $V^t = 0.1948$
	Low V^t	High V^t	Low V^t	High V^t

quality of the restaurant and the degree of the personal taste match. These two subsequent questions help us verify the driving forces of the purchase decision.[1]

The order of the quality component question and the personal taste component question was randomized among the respondents, so some were presented in the quality-taste order while others in the taste-quality order. The set of questions were specific for a given restaurant, and followed the manipulated reviews for each restaurant on the same computer screen.

[1]The responses to the two subsequent questions help us interpret the results. The answers to the quality component question in the main survey are consistent with those to the quality variance question in the pretest survey. The answers to the personal taste match question reveal that on average, respondents have stronger taste matchings for restaurant [2] compared to [1], and likewise for restaurant [7] compared to [8], indicating that personal taste matches are indeed explaining the purchase intent observed in [1] $<$ [2] and [7] $>$ [8].

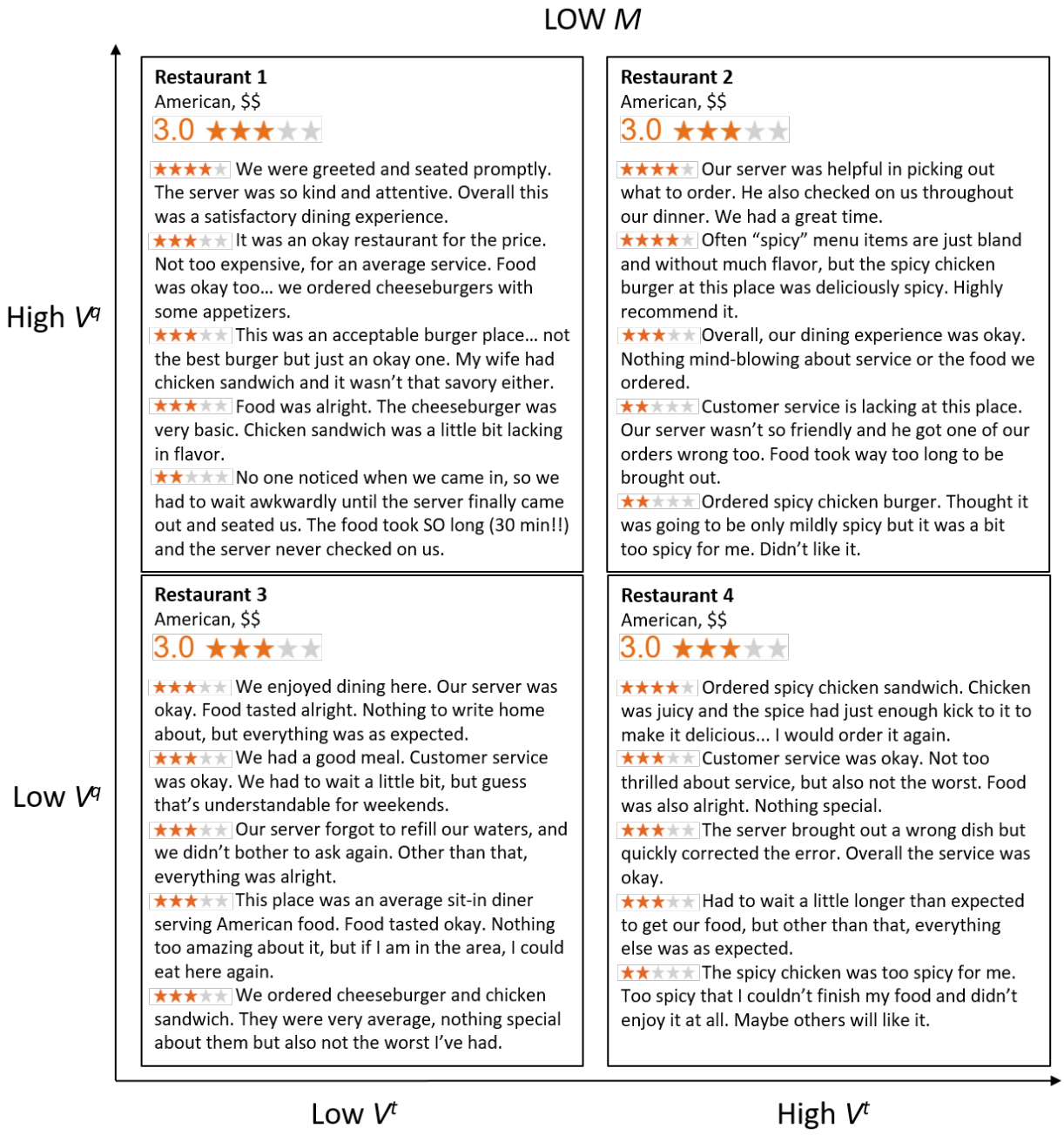

Figure OC.1: Manipulated Reviews for Low Mean Restaurants

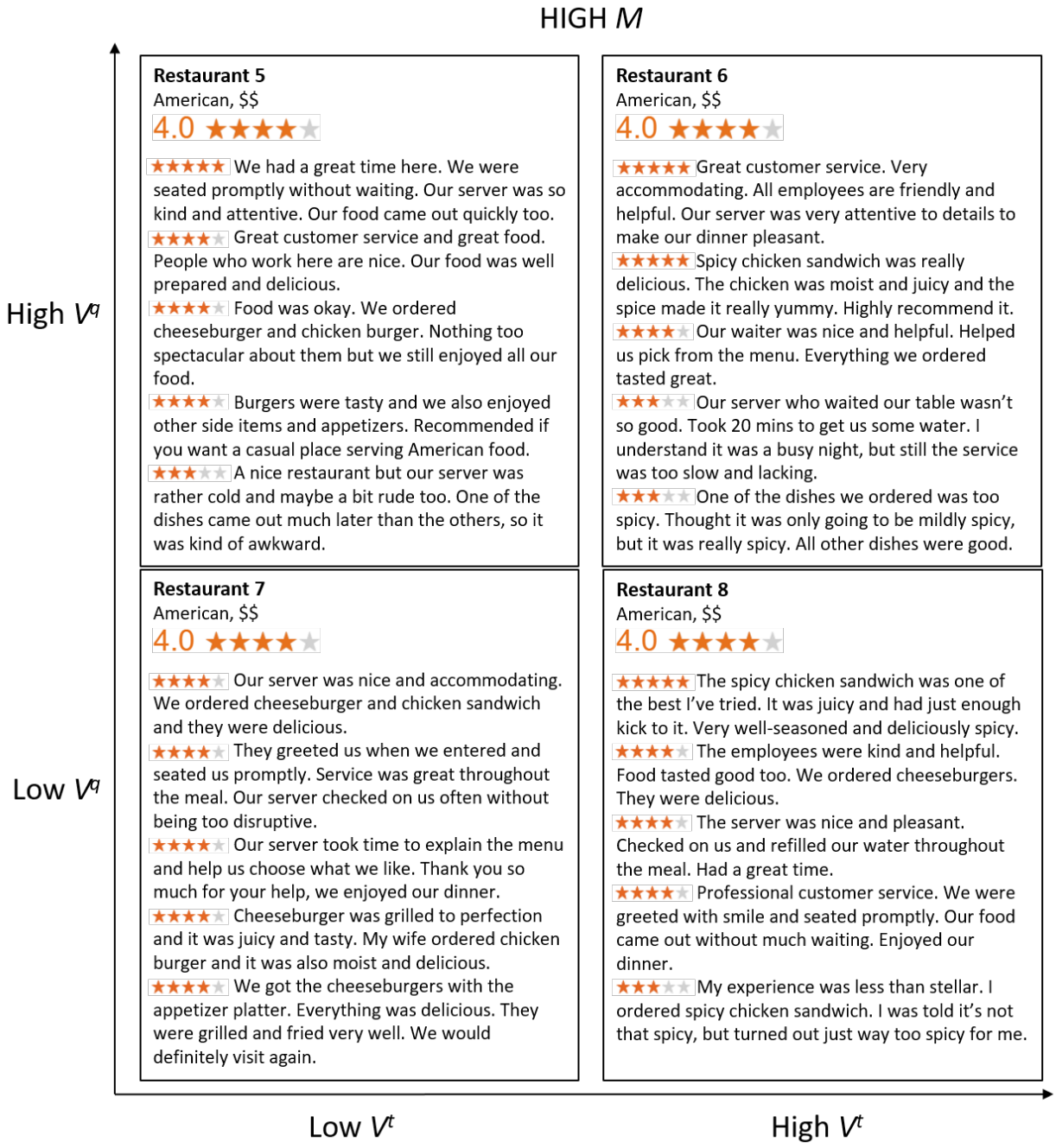

Figure OC.2: Manipulated Reviews for High Mean Restaurants

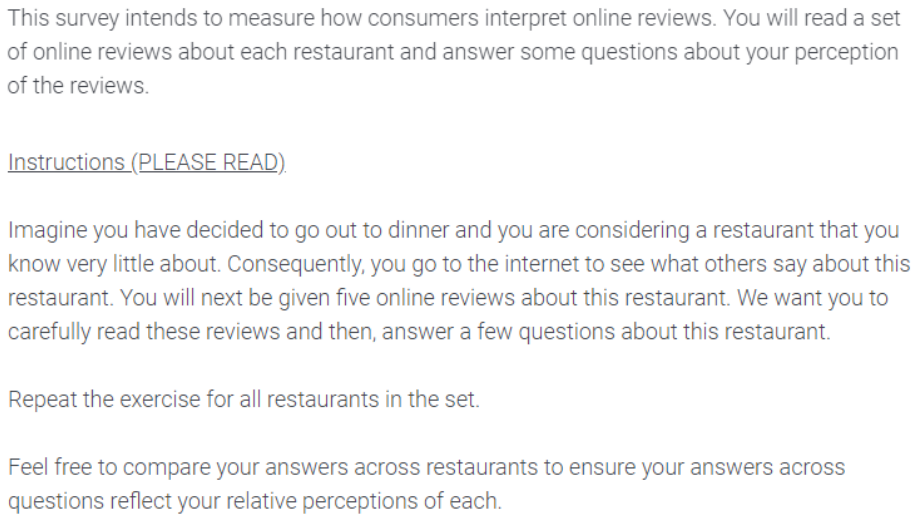

Figure OC.3: Instructions for the Pretest Survey

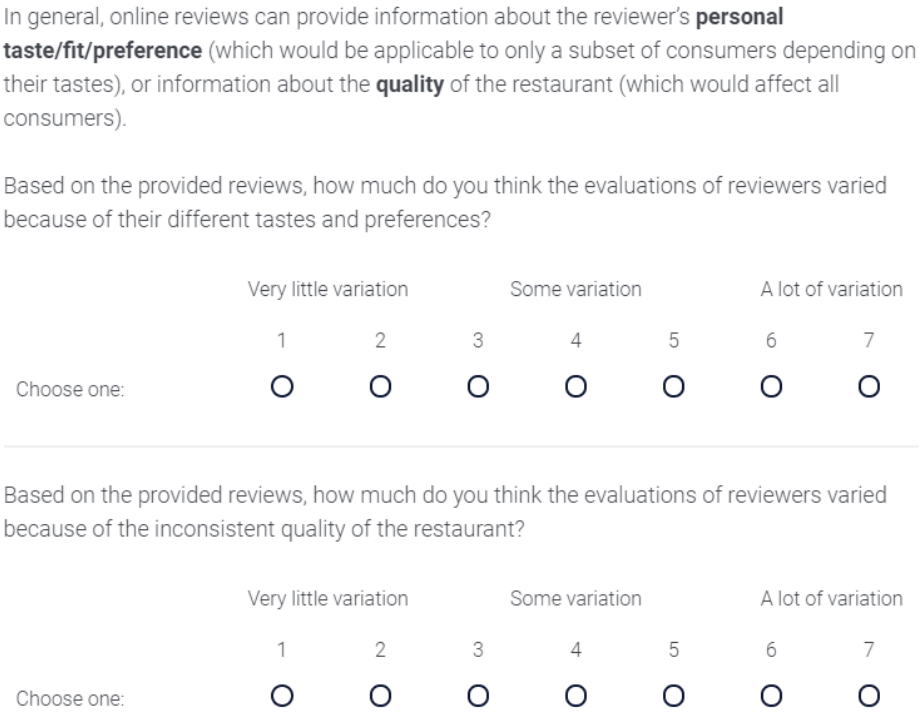

Figure OC.4: Pretest Survey Questions (Taste-Quality Order)

This survey intends to measure how consumers interpret online reviews. You will read a set of online reviews about each restaurant and answer some questions about your perception of the restaurant. The task should take about 10 minutes to complete.

All of your responses are completely confidential, so you may answer freely and openly.

Instruction (PLEASE READ)

You plan to eat at an American restaurant by yourself. There are 8 new American cuisine restaurants in your area you have yet to visit, so you read reviews on them. Based on the reviews on each restaurant (see below), please answer the questions regarding your perception of that restaurant.

Feel free to compare your answers across restaurants to ensure your answers across questions reflect your relative perceptions of each.

Figure OC.5: Instructions for the Main Survey

How likely are you (personally) to choose this restaurant?

	Not likely at all			Neutral			Highly likely
	1	2	3	4	5	6	7
Choose one:	○	○	○	○	○	○	○

In general, online reviews can provide information about the **quality** of the restaurant (which would affect all consumers), or information about the reviewer's **personal taste/fit/preference** (which would be applicable to only a subset of consumers depending on their tastes).

After reading the provided reviews, how consistent do you think the quality of this restaurant is?

	Not very consistent			Somewhat consistent			Highly consistent
	1	2	3	4	5	6	7
Choose one:	○	○	○	○	○	○	○

After reading the provided reviews, how much do you think the dining experience at this restaurant matches your own tastes and personal preferences?

	Very little			Somewhat			A lot
	1	2	3	4	5	6	7
Choose one:	○	○	○	○	○	○	○

Figure OC.6: Main Survey Questions (Quality-Taste Order)

OD Comparison of Lab Experiment Participants who Passed vs. Failed the Attention Check

While we presented the cell means and the hypotheses test results for the sample of participants who passed the attention check in the main paper, here we include those for the participants who failed the attention check (N=88) as a comparison (see subsection OD.1). Note that the attention check was presented as posing for another restaurant with the reviews and the set of questions in the same exact format as the actual questions in the survey, with just one out of the five text reviews containing a specific instruction to select a set of answers, as part of that one text review. This attention check question was placed at the end of the survey and was a very conservative test to check whether subjects were reading each and every review carefully until the very last question. Also note that the mean rating was clearly presented to the subjects, hence the attention to the mean rating of each restaurant is not part of what the attention check was intended to screen. The ratings for the five reviews were also visually presented to the respondents, and one could have figured out the total amount of variance in the ratings by visualizing the diagrams of the set of the star ratings. We also present the cell means and the hypotheses test results for the total sample, including both the subjects who passed and failed the attention check in subsection OD.2. From this set of analyses, we conclude that: 1) contrasts were much weaker and sometimes lacked significance for the subjects who failed the attention check, 2) increasing the number of respondents had an important role in finding differences in cell means, hence we found that our hypotheses were still supported when we included the entire set of respondents (N=178) (in other words, we were NOT cherry-picking the good answers and it would be fair to say that even those subjects who failed the check were giving the survey some level of attention), but 3) we would no longer see all of our hypotheses hold if we were to only use those subjects who failed the attention check. Therefore, our results were robust to whether we only use the subjects who passed the attention check (N=90) or the entire set of respondents (N=178), but we clearly see that the subjects who failed the attention check gave different responses compared to those who passed the check. Since we have no reason to rely ONLY on the set of respondents who failed the attention check, we conclude from

this robustness check that our set of hypotheses tests are supported in this lab experiment result.

OD.1 Respondents who Failed the Attention Check

The sample size of respondents is N=88 for all cells (restaurants).

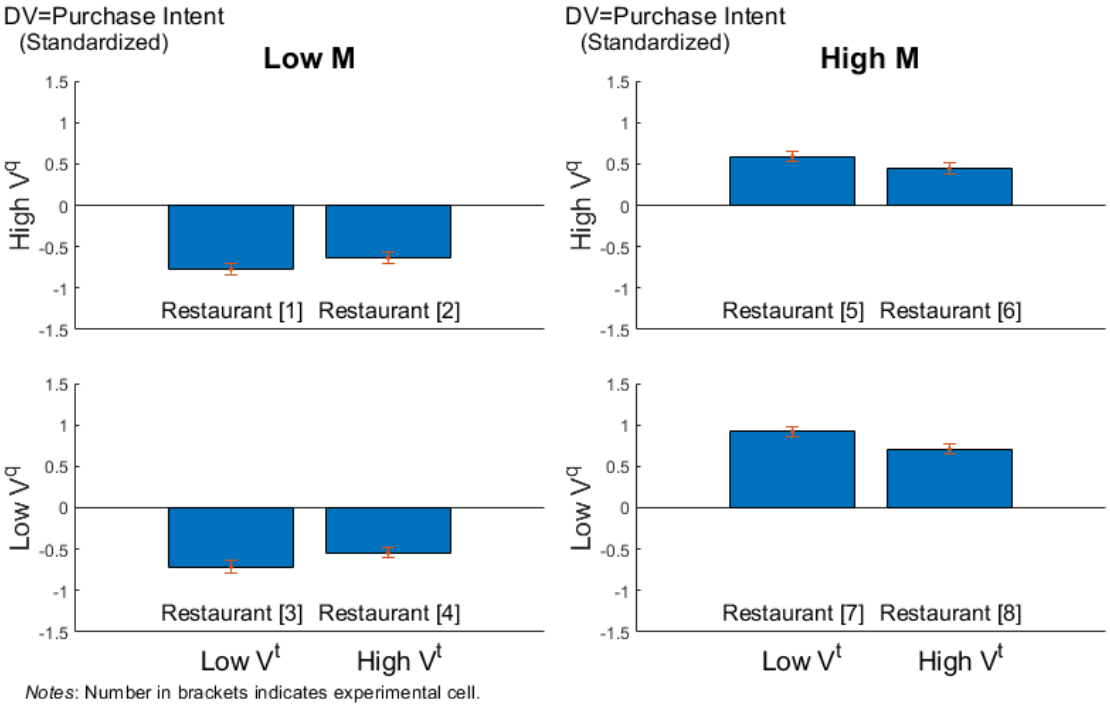

Figure OD.1: Purchase Intent from Respondents who Failed the AC

Table OD.1: Multiple Hypothesis Testing for Respondents who Failed the AC

Alternative Hypothesis on Purchase Intent		Raw p-value	Adjusted p-value
Effect of M is positive Restaurants [1],[2],[3],[4] < [5],[6],[7],[8] (pooled)		<0.0001	<0.0001
Effect of V^q is negative Restaurants [1],[2],[5],[6] < [3],[4],[7],[8] (pooled)		<0.0001	0.0002
Effect of V^t is more positive compared to high M & low V^q:	When M is low: Restaurants [8]−[7] < [4]−[3]	0.0020	0.0095
	When V^q is high: Restaurants [8]−[7] < [6]−[5]	0.2990	0.7689
	When M is low & V^q is high: Restaurants [8]−[7] < [2]−[1]	0.0041	0.0195

Notes: Number in brackets indicates experimental cell.

OD.2 Total Sample, Including the Respondents who Passed and Failed the Attention Check

The sample size of respondents is N=178 for all cells (restaurants).

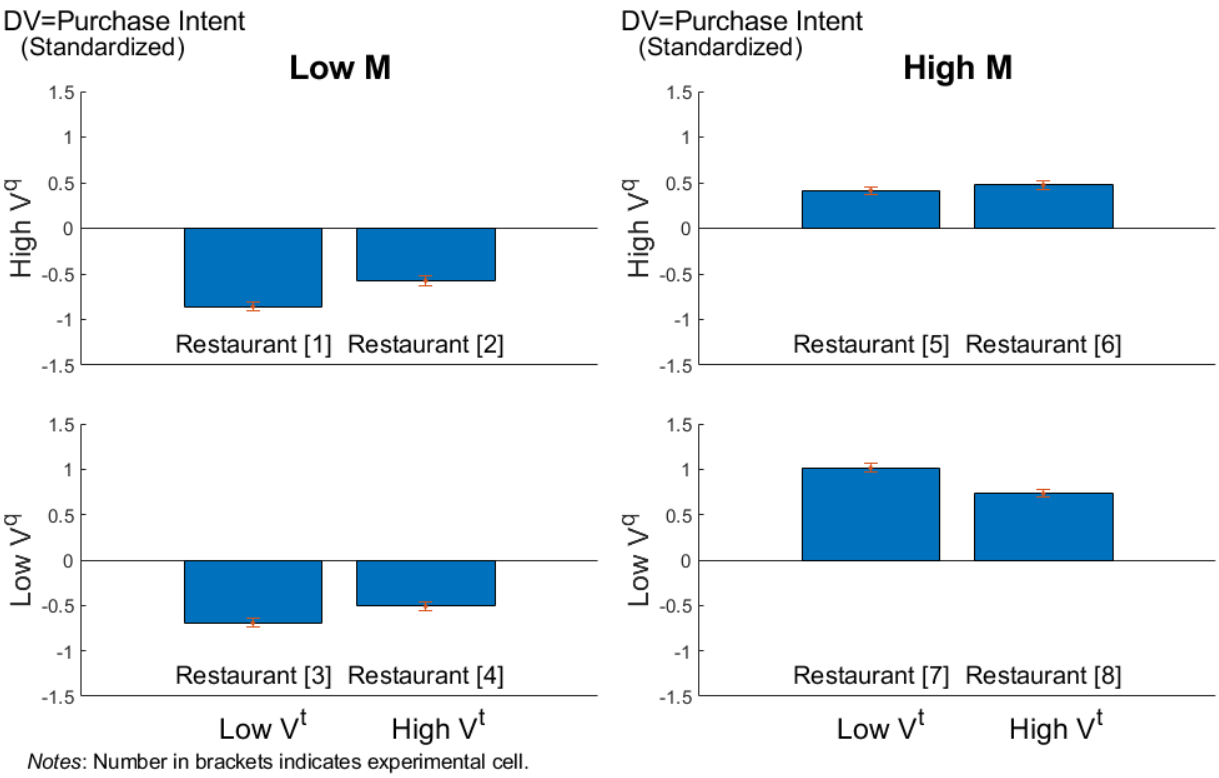

Figure OD.2: Purchase Intent from the Total Sample of Respondents

Table OD.2: Multiple Hypothesis Testing for the Total Sample of Respondents

Alternative Hypothesis on Purchase Intent		Raw p-value	Adjusted p-value
Effect of M is positive Restaurants [1],[2],[3],[4] < [5],[6],[7],[8] (pooled)		<0.0001	<0.0001
Effect of V^q is negative Restaurants [1],[2],[5],[6] < [3],[4],[7],[8] (pooled)		<0.0001	<0.0001
Effect of V^t is more positive compared to high M & low V^q:	When M is low: Restaurants [8]−[7] < [4]−[3]	<0.0001	<0.0001
	When V^q is high: Restaurants [8]−[7] < [6]−[5]	0.0001	0.0008
	When M is low & V^q is high: Restaurants [8]−[7] < [2]−[1]	<0.0001	<0.0001

Notes: Number in brackets indicates experimental cell.

OE NLP Algorithm

We implemented the Natural Language Processing (NLP) algorithm as follows. First, we transformed review texts into a vector of words. All texts underwent the same algorithm in this step (and this was also used in picking out 5,000 samples for surveys). We corrected word elongation that is commonly present in online reviews (e.g., "awwwwwesome" to "awesome"). We also converted common internet slangs to normal languages ("TGIF" to "Thank God it's Friday"). Three or more capitalized characters with spaces in-between typically denote emphasized words, and they are put together ("W O R L D" to "WORLD"). We also eliminated common first and last names (as names have less to do with quality versus personal taste issues in reviews) before feeding the texts into the spell checker.

The spell checker compares each word in the text to those in a standard dictionary; if the word is not found, the spell checker attempts to correct the word to one found in the dictionary. The algorithm identifies the word with lower edit distance but gives preference to frequently occurring words. When the spell corrector cannot fix the misspelled word, it will attempt to break the term into two parts, as sometimes simply a space is missing between two separate words. If unsuccessful, the misspelled word is left as is. The spell correction step is crucial because crowd-sourced text content from the web are inundated with misspellings and typos. For example, without implementing the spell corrector, the entire set of 283,069 restaurant reviews would have contained 65,997 words (instead of 28,572 words after the spell correction). This is problematic not only because it increases the number of predictors but more importantly because it does not count the word frequencies accurately. Occasionally, the spell checker creates false corrections, but the gain is expected to be far greater.

After the spell corrector was applied, we transformed all the text to lower case. Then we removed "stop words", such as "a", "is", "the", etc. We converted the dollar sign ("$") to the word "dollars", before removing punctuation. We also removed numbers; often reviews would discuss specific price ranges, but it is not the numerical characters but the words surrounding them ("dollars", "high price", "expensive") that are indicative of the discussion. After all these transformations, the words were stemmed into their root form, and all texts were cast into a frequency vector of rooted words.

OF Infogroup's Estimated Sales

A conversation with an Infogroup representative revealed that the company estimates the sales figures using the following methodology: total sales and the number of the employees on payroll for each industry are published by the Economic Census every five years (the recent years being 2007, 2012, and 2017). From this, Infogroup gathers sales per employee for each industry and county, which serve as the reference points to complement the location-specific employee numbers that Infogroup gathers first-hand through surveys and phone calls. Then, using a proprietary model that incorporates the number of years in operation and price changes for each industry (available from the Bureau of Economic Analysis), the company estimates the annual sales per each business location. Hence, the variables that are used to estimate the sales include: county-specific and industry-specific references for sales per employee, the number of employees at each specific business location, the number of years in operation, and the industry norms of the Department of Commerce, such as price changes and year effects.

We cross-checked the Infogroup's estimated sales for Texas hotels against the actual revenues we obtained from the governmental tax agency; the scatterplot of the two revenues display a positive correlation (.8) and is shown in Figure OF.1.

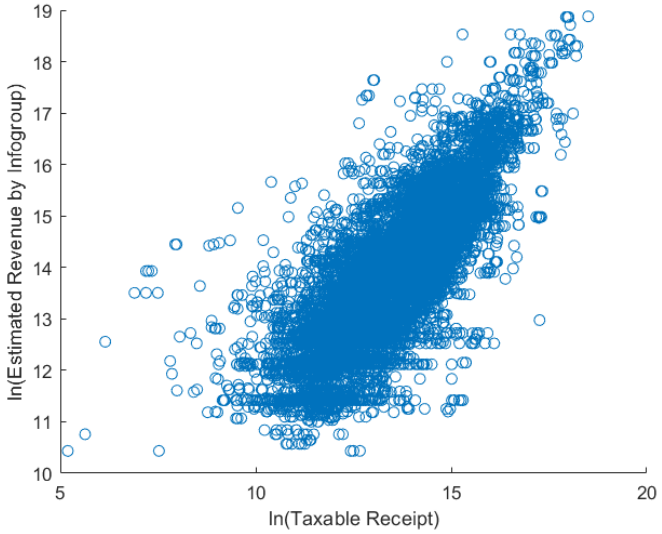

Figure OF.1: Comparison of Infogroup's Estimated Revenue vs. Government-reported Revenue for Texas Hotels

Furthermore, we checked that the measurement error for the Texas hotel data was not systematically associated with our review variables of interest, by running the following OLS regression with the measurement error as the DV:

$$(\text{estimated sales} - \text{actual sales}) = aM + bV^q + cV^t + [\text{policy, firm and time dummies}] + \text{error}. \quad \text{(OF.1)}$$

None of the relevant coefficients, i.e., a, b and c, was even marginally significant. We report the regression results in Table OF.1, with the comparison output when using the actual revenue as the DV. Although there is no way to be certain that the same will hold for California restaurants as well, we conclude from this analysis that these insights provide strong supporting evidence that any measurement error associated with Infogroup's methodology is not systematically associated with our review variables, a condition that is needed to render our restaurant field results as unbiased.

Table OF.1: Infogroup Measurement Error OLS Results

	DV: Actual Revenue				DV: Measurement Error			
	Coef.	Std. Err.	t	p-value	Coef.	Std. Err.	t	p-value
M	0.0178	(0.0073)	2.45	0.014**	-0.0146	(0.0110)	-1.32	0.188
V^q	-0.0141	(0.0073)	-1.95	0.052*	0.0107	(0.0111)	0.97	0.333
V^t	0.0076	(0.0075)	1.02	0.309	0.0038	(0.0114)	0.33	0.742
Pol	0.0129	(0.0080)	1.61	0.108	0.0104	(0.0122)	0.85	0.395
Firm Fixed Effects	Included				Included			
Time Fixed Effects	Included				Included			
Num Obs.	10983				10956			
Num Panelists	2359				2359			
Avg Obs./Panelist	4.7 (years)				4.6 (years)			

***: significant at 1%, **: significant at 5%, and *: significant at 10% (**two-tailed tests**).

OG Matching Businesses Across Two Databases

Given that Infogroup is not a commonly used database (at least in the marketing literature), we explain the algorithm used to match a business entry in the Infogroup's database with that on the Google business listing. Two different databases may follow different conventions for business names and addresses making it difficult to match a firm across the two different databases. An illustrative example from Chaudhuri et al. (2003) highlights the difficulty in making the correct matchings; Among three reference companies "Boeing Company", "Bon Corporation", and "Companions", there is a high probability of misclassification of "Boeing Corporation" to "Bon Corporation", and "Company Boeing" to "Companions" if we are not careful. In such a case, words have to be given different importance based on the frequency of occurrences in the database (so that unique words such as "Boeing" are given more importance). A fuzzy matching algorithm may work best in a broad setting, although some threshold will have to be set for the acceptance of the matching, so there is a trade-off between false matches and missed matches. For our analysis, we prefer having a smaller set of firms with some omitted matches over having too many false matches in our end result dataset. We note that in our setting, the task is immensely simplified with the inclusion of geo-spatial coordinate variables. Both the Infogroup and Google's database contain latitude and longitude for each business, and the quality of these variables is exceptional in that there was 100% fill-rate in both databases, and the discrepancy in coordinates across databases is typically on a scale of .0001 degrees. Using the fact that a local business establishment has to occupy the same physical location, we greatly reduce our search scope by only looking at the candidate businesses within .002 degrees of coordinates in the other database. Among those, then, we have the business name and the address information to make the correct matchings across databases.

Upon examining both databases' name and address conventions, we find that the discrepancies typically exist at the end of the string variable for both the business name and the address. For names, we may have "SUBWAY", "SUBWAY SANDWICHES", or "SUBWAY CO.", but fortunately neither database would have "SANDWICH SUBWAY". For addresses, we may have "123 ANDERSON ST." or "123 ANDERSON STREET, STE 102", but the

street number and the street address are accurate in both databases. Therefore, we utilize the characteristics of our databases in order to sort through the entries in both databases in a relatively efficient way. After identifying the set of businesses within .002 degrees distance in latitude/longitude (which is about 200 meters) from the reference business, we search for the entry that has the identical 5 first characters in business name and the identical 10 first characters in business address. If the search returns a one-to-one matching, we accept the assignment (this occurs for a little over half of the successful first-run matches). If the reference business (from Google) is matched with more than two entries in the Infogroup database with different business names and all other identifiers (so they are indeed different listings in the Infogroup database), then we need to make a selection. This occurs when the first word of the business name is not unique enough and those multiple businesses are housed in the same building with identical address (e.g. "DALY CITY DINERS" and "DALY CITY BREWING COMPANY" located in the same building). In such a case, we calculate the edit distance and choose the one with the minimum edit distance in business names. Edit distance is a dynamic programming problem that finds the minimum required number of character edit operations required to transform one string vector into another (Navarro, 2001; Navarro et al., 2001). With our much trimmed-down number of businesses to choose from, the simple edit distance criterion is good enough to make the correct matching.

Out of 7,663 restaurants listed on Google in the San Francisco area, 4,305 (56%) were successfully matched with an entry in the Infogroup database using our methodology. The matched and unmatched restaurants had similar values of review statistics, as well as a comparable distribution of the number of reviews.

We used a similar approach to match the taxpayer's data from the Texas Comptroller's office to the hotel listings on Google reviews and to Infogroup's database in the hotel industry.

OH Simulation

To better assess our estimation approach and underlying model assumptions, and to address any concerns that we used estimated revenues for restaurants, we run a simulation using setups similar to the one in our field studies and provide estimates from the simulated data, along with a number of sensitivity analyses, in this appendix. The goal of this simulation exercise is three-fold. The first is to see if we can obtain reasonable estimates of the size and the sign of the marginal effects of our three observed factors, i.e., the mean, the quality variance, and the taste variance of the ratings, on subsequent demand. The second is to determine if the generated data is "rich" enough to allow us to get the hypothesized effects using our IV approach for at least one set of generated data. Third, we want to broaden this exploration to show that for a wide range of environments we can successfully recover the hypothesized relationships put forth in our model development. Before addressing these three goals, we first present a verbal description of the simulation and follow this with the actual code. We then present our analyses.

We assume for our base case a risk coefficient of one ($\alpha = 1$) and simulate quality and taste mismatch realizations using 2,000 firms over 20 time periods. We also assume an initial expected demand for each establishment to be 20.[2] At the beginning of the program, for each firm the initial demand is drawn from a Poisson distribution with a parameter 20. This initial demand is (by assumption) the number of reviews that will be provided in the first period. Each firm has a (randomly drawn) time-invariant quality and a (randomly drawn) taste location. For the set of reviewers for the first period, their experiences are also randomly determined based in part on the firm's quality level. We do this by randomly drawing from a uniform distribution a quality shock and a taste location for each reviewer. Additionally, we also randomly draw the reviewer harshness, again using a uniform distribution. The rating given by each reviewer is determined by 1) the quality realization, 2) the taste mismatch cost, i.e., the distance between the reviewer's location and the firm's location in the taste dimension, and 3) the reviewer harshness.

From the set of reviews left by the first period reviewers, we need to calculate the review

[2]We will show subsequently that our estimates are robust to different ranges of values for these settings.

statistics, i.e., the mean, the quality variance, and the taste variance of the ratings, as utilized by consumers reading these reviews. In order to do so, we "modify" the actual experiences by the reviewer's harshness, so the experiences reflected in the ratings are influenced by the reviewer harshness as well as the true experiences. Specifically, we make the following assumptions, in line with the assumptions made in the main paper. First, we determine the percent of each review devoted to quality based on our process model of how reviewers determine what to write about, i.e., we calculate the ratio of the absolute value of the true quality shock divided by the sum of the absolute value of both types of shocks. Then, we assume that the reviewer harshness is proportionately distributed across quality and taste evaluations (i.e., a harsh reviewer is harsh in evaluating both the quality and the taste aspects of the product). Note that because we are simulating the underlying data, we actually know the true quality and taste shocks, but we are "pretending" that the reviewer harshness is mixed in with these quality and taste shocks so that we can calculate the quality and taste variances incorporating these reviewer harshness, as observed by an outside observer. This mimics the real world setting in our field study, and allows us to estimate the IV results in addition to the OLS estimates.

Once the reviewer harshness is appropriately attributed to quality and taste dimensions, we calculate the review statistics for these first period reviews as well as the instrumental variables from the reviewer harshness for the same time period. Then, we calculate the expected demand (i.e., the Poisson parameter) using the second period demand function from our analytical model, i.e., Equation 9 in the paper. Hence, the predicted demand in the following period is a function of the observed review mean, the quality variance, and the taste variance in the current period (which reflect the true experiences as well as the reviewer harshness). Using the predicted demand as the expected demand, we randomly draw the next period's demand from a Poisson distribution. As before, this next period's demand represents the number of reviews that will be written in the following period, and the program continues until the final period, for all firms. The specific functional forms of the quality and taste shocks are provided within our simulation program, which we provide here in full:

```
%Setup parameters
alpha = 1; %risk coefficient
lambda0 = 20; %initial expected demand
J = 2000; %number of firms
T = 20; %number of periods

%Initialization
M = zeros(J*T,1);
Vq = zeros(J*T,1);
Vt = zeros(J*T,1);
D = zeros(J*T,1);

for j=1:J

   nreviews = poissrnd(lambda0); %draw initial demand from Poisson(lambda0)
   rest_timeinv_harsh = rand; %draw restaurant time-invariant harshness from
       Uniform(0,1)
   rest_timeinv_qual = rand; %draw restaurant time-invariant quality
   taste_j = 3*rand; %draw restaurant taste location from 3*Uniform(0,1)

   for t=1:T

      reviewer_harsh = 2*rand(1,nreviews); %draw a vector of reviewer harshness
      harshness = -reviewer_harsh - rest_timeinv_harsh; %set harshness
          realizations
      quality = 4 + rest_timeinv_qual + rand(1,nreviews); %draw quality shocks
          and set quality realizations
      taste_i = 3*rand(1,nreviews); %draw reviewers' taste locations

      qualdev = quality - 3 - rest_timeinv_qual; %set quality deviations as the
          randomly drawn quality shocks + 1
      tastedev = taste_i - taste_j; %calculate taste deviations
      q = abs(qualdev)./(abs(qualdev)+abs(tastedev)); %calcualte the fraction of
          quality deviations relative to the total deviations

      reviews = harshness + quality - abs(taste_i - taste_j); %calculate the
          ratings
      reviews = min(5,max(1,reviews)); %cap ratings from 1 to 5; not absolutely
          mandatory
```

```
    %Calculate the review statistics for firm j, period t
    M((j-1)*T+t,1) = mean(reviews);
    Vq((j-1)*T+t,1) = var(q.*harshness + quality);
    Vt((j-1)*T+t,1) = var((1-q).*harshness - abs(taste_i - taste_j));

    %Calculate the instruments
    mean_harsh((j-1)*T+t,1) = mean(harshness);
    varq_harsh((j-1)*T+t,1) = var(q.*harshness);
    vart_harsh((j-1)*T+t,1) = var((1-q).*harshness);

    %Calculate demand for the next period
    lambda = nreviews/4 .* (M((j-1)*T+t,1) + sqrt(3.*Vq((j-1)*T+t,1)) +
        sqrt(3.*Vt((j-1)*T+t,1)) - 1./alpha .*
        (log(1/(2*alpha*sqrt(3*Vq((j-1)*T+t,1)))) +
        log(exp(2*alpha*sqrt(3*Vq((j-1)*T+t,1)))-1))./sqrt(3*Vt((j-1)*T+t,1));
        %demand prediction from the analytical model
    lambda = max(lambda,0); %keep the expected demand positive
    D((j-1)*T+t,1) = poissrnd(lambda); %draw next period's demand
    D = min(100,max(2,D)); %cap demand from 2 to 100
    nreviews = D((j-1)*T+t,1); %set next period's number of reviews

  end
end
```

Using this setup, we generate one simulation for this base case and obtain a distribution of ratings as shown in Figure OH.1, and the scatterplots of the review mean against the two types of the review variance are shown in Figures OH.2 and OH.3. As can be seen from these figures the mean ratings are approximately trapezoidal for this base case with a mean about 2.5 and the two variances are (as expected) larger when the mean is about average, i.e., 2.5, and lower for firms that have high and low means (due to the ceiling effect).

We then use this base case simulated data to address our three questions. We first investigate the appropriateness of our approach by obtaining "rough" predictions of the marginal effects of the mean and the two variances on subsequent demand in order to compare them with our empirically estimated marginal effects. We say "rough" since the comparative static results are for a specific firm with specific values of the mean and the two variances,

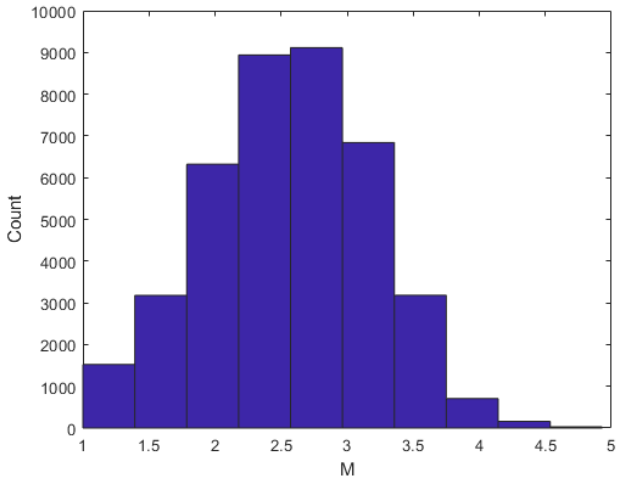

Figure OH.1: Distribution of the Mean of the Simulated Ratings

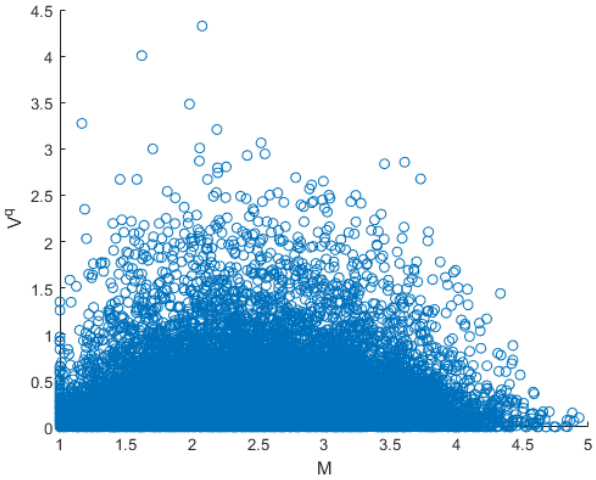

Figure OH.2: Scatterplot of Mean vs. Quality Variance of Simulated Ratings

yet our empirical results are based on pooling the data from all firms and estimating the best fitting marginal effects. Moreover, we are assuming linear effects even though we know the demand function is non-linear and thus, as seen below, these marginal effects will vary somewhat depending on where the firm is on the demand surface.

$$\frac{\partial D_2^*}{\partial M} = \frac{D_1}{4\sqrt{3V_t}}; \tag{OH.1}$$

$$\frac{\partial D_2^*}{\partial V_q} = \frac{D_1}{8\sqrt{V_q V_t}} \cdot \left(\frac{1}{\alpha\sqrt{3V_q}} + \frac{(1+e^{2\alpha\sqrt{3V_q}})}{(1-e^{2\alpha\sqrt{3V_q}})} \right); \quad \text{and} \tag{OH.2}$$

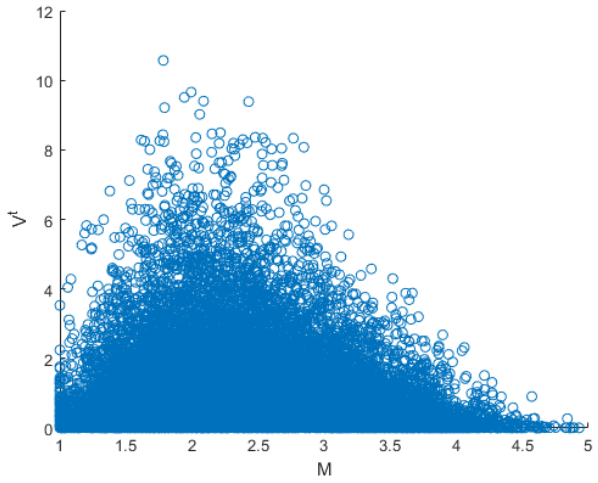

Figure OH.3: Scatterplot of Mean vs. Taste Variance of Simulated Ratings

$$\frac{\partial D_2^*}{\partial V_t} = \frac{D_1}{8V_t} \cdot \left(1 - \frac{1}{\sqrt{3V_t}} \cdot \left(M + \sqrt{3V_q} + \sqrt{3V_t} - \frac{1}{\alpha}\left(ln\frac{1}{2\alpha\sqrt{3V_q}} + ln(e^{2\alpha\sqrt{3V_q}} - 1)\right)\right)\right). \tag{OH.3}$$

Given this caveat we obtain these estimations as follows. For each observation (i.e., firm per period) in our simulated data, there is an associated realized values for M, V^q, and V^t. We can use these values to calculate the above comparative statics to obtain the predicted marginal effect of M, V^q, and V^t.[3] These rough analytical estimates are reported in the first column of Table OH.1. As expected, these four values have the right sign and will serve as a benchmark to compare our empirical estimates against.

We next use the same strategy delineated in the main paper to estimate the effects of the quality and the taste variances, and the interaction between the two, on subsequent demand. We use median cuts to divide the sample into high and low levels of the mean rating and the quality variance.[4] We report the OLS and the 2SLS estimation results for our "direct test" version of the specification in the second and third columns of Table OH.1, while the first-stage estimates for the 2SLS are found in Table OH.2. We find that the first-

[3]For the sample of our observations, we can find the average effect by calculating the average of all the marginal effects, but because these comparative statics are highly non-linear and our simulated data contain randomly drawn realizations, instead we report the predicted effect of an "average datum" in our sample (which is similar in spirit to what regression finds). Also note that the interaction term is obtained by calculating the predicted effect of an "average datum" within low mean and high quality variance conditions, and taking the difference between the base effect of an average datum across all observations.

[4]We do not have a skewed distribution of ratings in our simulated data, so we just use the median value of V^q to divide the sample into high and low V^q, instead of regressing V^q on M to find high and low values of V^q given the level of M, as was done in our field setting with a skewed distribution of ratings.

Table OH.1: Simulation Estimation Results - Direct Test of Hypotheses

				OLS			IV(2SLS)		
Demand	Prediction:		Analytical¶:	Coef.	Std. Err.	p-value	Coef.	Std. Err.	p-value
M	γ_1	$\gamma_1 > 0$	1.1714	2.0593***	(0.0921)	3.9913e-110	1.5698***	(0.1608)	1.7364e-22
V^q	γ_2	$\gamma_2 < 0$	-0.5563	-0.8967***	(0.1560)	9.0717e-09	-0.6004**	(0.2340)	0.0103
V^t	γ_3	$\gamma_3 < 0$	-2.2901	-2.4025***	(0.0912)	1.0042e-151	-2.3921***	(0.1126)	1.5440e-99
$V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$	γ_7	$\gamma_7 > 0$	0.9016	1.2798***	(0.0943)	7.6374e-42	1.1842***	(0.1207)	1.0321e-22
Firm Fixed Effects				Included			Included		
Num Obs.				40000			40000		
Num Panelists				2000			2000		
Obs./Panelist				20			20		
R^2:				0.0357			0.0181		
F:				370	(p-val=1.43e-313)		185	(p-val=3.84e-157)	

***: significant at 1%, **: significant at 5%, and *: significant at 10% (**two-tailed tests**).

¶: Analytical calculations are rough approximations for our simulated data.

stage identification is strong enough (in our simulation, we ensured this to be the condition), and in the second-stage we recover the expected signs for the effects of our explanatory variables. The size of the effects in the 2SLS estimates are also comparable to that in the OLS regression, albeit smaller (as expected).

Next, we compare the calculated figures based on the comparative statics with our OLS estimates. Although they are not a perfect match, they are similar to (but happen to be smaller than) our OLS estimates and show the same signs. Although much of this comparison should be anticipated (since we used our demand prediction from the analytical model to cast the next period's demand within our simulation, and the comparative statics are from the same demand equation), we would not expect exact matches due to the fact that we are not estimating each firm's non-linear demand function, but instead a pooled demand across observations. Moreover, our simulation varies somewhat from our analytic model, i.e., randomness is added at various draws at each stage of the simulation, and some settings vary slightly from our theoretical setup (e.g., in the simulation, we have a wider support of consumers who would consider a restaurant, the demand and the ratings are censored, etc.). Nonetheless, the direction of the effect of each of the review variables should be preserved,

Table OH.2: First-Stage Estimates

	Direct Test of Hypotheses											
	DV: M			DV: V^q			DV: V^t			DV: $V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$		
	Coef.	Std. Err.	p-value	Coef.	Std. Err.	p-value	Coef.	Std. Err.	p-value	Coef.	Std. Err.	p-value
User M	0.9179	(0.0062)	0***	0.0220	(0.0033)	2.4233e-11***	0.2375	(0.0094)	4.1933e-140***	0.0806	(0.0088)	5.6256e-20***
User V^q	0.3296	(0.0083)	0***	0.8784	(0.0044)	0***	0.7160	(0.0125)	0***	0.7376	(0.0117)	0***
User V^t	0.7975	(0.0312)	5.2713e-143***	-0.4059	(0.0165)	1.3808e-132***	6.2086	(0.0470)	0***	-0.6859	(0.0440)	1.5165e-54***
User $V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$	-0.9874	(0.0314)	2.7021e-214***	0.4351	(0.0166)	7.6743e-150***	-1.0185	(0.0473)	2.4673e-102***	5.9236	(0.0443)	0***
Firm Fixed Effects	Included			Included			Included			Included		
Num Obs.	40000			40000			40000			40000		
Num Panelists	2000			2000			2000			2000		
Obs./Panelist	20			20			20			20		
R^2:	0.399			0.520			0.611			0.657		
F:	6.63e+03	(p-val=0)		1.08e+04	(p-val=0)		1.57e+04	(p-val=0)		1.92e+04	(p-val=0)	

***: significant at 1%, **: significant at 5%, and *: significant at 10% (**two-tailed tests**).

Table OH.3: Simulation Estimation Results - Average of 20 Simulations

$\alpha = 1$, $D_0 = 20$, $J = 2000$, $T = 20$				
	OLS		IV(2SLS)	
Demand	Coef.	Std. Err.	Coef.	Std. Err.
M	2.0975	(0.0381)	1.6656	(0.0434)
V^q	-0.8131	(0.0389)	-0.6133	(0.0570)
V^t	-2.3279	(0.0259)	-2.2798	(0.0259)
$V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$	1.2794	(0.0236)	1.1620	(0.0254)
Firm Fixed Effects	Included		Included	

as predicted by the comparative statics equations, and we confirm this using our rough calculations on our dataset.

The above analyses are for one run of the simulation using our base case settings. We check for the stability of these results by repeating this base case simulation 19 more times using different random number seeds and provide in Table OH.3 the average values of the coefficients along with the standard errors (the standard error is across the 20 simulations, i.e., it is the standard deviation among the 20 coefficients, divided by the square root of 20; hence, the standard errors capture how stable the coefficients are across simulations). As can be seen from this table, the individual simulation runs produce very similar results.

In summary, these results show that our estimation approach is able to recover the main and "interaction" effects that our analytic model would predict at least for one base case setting.

We next explore the robustness of these results by varying four different parameters assumed in the base case, these being the risk coefficient, the initial expected demand, the number of firms and periods. For each analysis, we run the simulation 20 times and we present these average coefficients across the relevant 20 simulations in Tables OH.4 - OH.11. As can be seen from these tables, we find robust results in terms of significance and sign, regardless of the assumed parameter.

Table OH.4: Sensitivity Analysis - Lower Risk Coefficient

$\alpha = 0.5$, $D_0 = 20$, $J = 2000$, $T = 20$				
	OLS		IV(2SLS)	
Demand	Coef.	Std. Err.	Coef.	Std. Err.
M	2.0503	(0.0271)	1.5566	(0.0373)
V^q	-0.6717	(0.0282)	-0.4893	(0.0336)
V^t	-2.3321	(0.0252)	-2.2764	(0.0231)
$V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$	1.2400	(0.0229)	1.1119	(0.0256)
Firm Fixed Effects	Included		Included	

Table OH.5: Sensitivity Analysis - Higher Risk Coefficient

$\alpha = 2$, $D_0 = 20$, $J = 2000$, $T = 20$				
	OLS		IV(2SLS)	
Demand	Coef.	Std. Err.	Coef.	Std. Err.
M	2.0112	(0.0240)	1.5230	(0.0465)
V^q	-1.0179	(0.0287)	-0.7138	(0.0476)
V^t	-2.2877	(0.0202)	-2.2194	(0.0243)
$V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$	1.2767	(0.0166)	1.1164	(0.0263)
Firm Fixed Effects	Included		Included	

Table OH.6: Sensitivity Analysis - Lower Initial Demand

$\alpha = 1$, $D_0 = 10$, $J = 2000$, $T = 20$				
	OLS		IV(2SLS)	
Demand	Coef.	Std. Err.	Coef.	Std. Err.
M	1.9452	(0.0381)	1.4386	(0.0551)
V^q	-0.7203	(0.0508)	-0.4781	(0.0605)
V^t	-2.0855	(0.0819)	-2.0565	(0.0768)
$V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$	0.9823	(0.0974)	0.8879	(0.0904)
Firm Fixed Effects	Included		Included	

Table OH.7: Sensitivity Analysis - Higher Initial Demand

$\alpha = 1$, $D_0 = 30$, $J = 2000$, $T = 20$				
	OLS		IV(2SLS)	
Demand	Coef.	Std. Err.	Coef.	Std. Err.
M	2.1284	(0.0185)	1.6851	(0.0303)
V^q	-0.9087	(0.0384)	-0.6289	(0.0421)
V^t	-2.4482	(0.0221)	-2.3828	(0.0249)
$V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$	1.4242	(0.0187)	1.2824	(0.0218)
Firm Fixed Effects	Included		Included	

Table OH.8: Sensitivity Analysis - Lower Number of Firms

$\alpha = 1$, $D_0 = 20$, $J = 1000$, $T = 20$				
	OLS		IV(2SLS)	
Demand	Coef.	Std. Err.	Coef.	Std. Err.
M	2.1020	(0.0505)	1.7343	(0.0537)
V^q	-0.8291	(0.0562)	-0.6398	(0.0872)
V^t	-2.3120	(0.0393)	-2.2733	(0.0430)
$V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$	1.2742	(0.0351)	1.1714	(0.0423)
Firm Fixed Effects	Included		Included	

Table OH.9: Sensitivity Analysis - Higher Number of Firms

$\alpha = 1$, $D_0 = 20$, $J = 4000$, $T = 20$				
	OLS		IV(2SLS)	
Demand	Coef.	Std. Err.	Coef.	Std. Err.
M	2.0883	(0.0299)	1.6194	(0.0344)
V^q	-0.8120	(0.0200)	-0.6190	(0.0328)
V^t	-2.3346	(0.0187)	-2.2759	(0.0199)
$V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$	1.2797	(0.0160)	1.1468	(0.0193)
Firm Fixed Effects	Included		Included	

Table OH.10: Sensitivity Analysis - Lower Number of Periods

$\alpha = 1$, $D_0 = 20$, $J = 2000$, $T = 10$				
	OLS		IV(2SLS)	
Demand	Coef.	Std. Err.	Coef.	Std. Err.
M	1.9928	(0.0451)	1.5345	(0.0444)
V^q	-0.7704	(0.0357)	-0.5667	(0.0715)
V^t	-2.1152	(0.0248)	-2.0339	(0.0308)
$V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$	1.0448	(0.0261)	0.8930	(0.0324)
Firm Fixed Effects	Included		Included	

Table OH.11: Sensitivity Analysis - Higher Number of Periods

$\alpha = 1$, $D_0 = 20$, $J = 2000$, $T = 50$				
	OLS		IV(2SLS)	
Demand	Coef.	Std. Err.	Coef.	Std. Err.
M	2.0949	(0.0154)	1.6341	(0.0275)
V^q	-0.8238	(0.0209)	-0.6712	(0.0226)
V^t	-2.4679	(0.0146)	-2.4514	(0.0175)
$V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$	1.4280	(0.0147)	1.3451	(0.0176)
Firm Fixed Effects	Included		Included	

OI First-Stage Estimates of 2SLS

We report the first-stage estimates of the 2SLS main estimation results. Table OI.1 contains the first-stage estimates for the 2x2 specification for the yearly restaurant panel (6 endogenous variables, one in each column). Table OI.2 is for the alternative specification for the direct test of the hypotheses for the restaurants (4 endogenous variables). Similarly, the first-stage estimates for the quarterly hotel panel are shown in Table OI.3 and OI.4 for each respective specification.

Note that the instrumental variables are denoted as "User M", "User V^q", etc., to indicate the corresponding statistics among the users' historic means. For the exact definitions of these IVs, please refer to the last two paragraphs of the estimation subsection in the main paper.

Table OI.1: First-Stage Estimates

	Restaurants (Yearly Panel), 2x2 Specification					
	DV: M	DV: V^q	DV: V^t	DV: $V^t \times \mathbb{1}_{\text{low}M \,\&\, \text{low}V^q}$	DV: $V^t \times \mathbb{1}_{\text{high}M \,\&\, \text{high}V^q}$	DV: $V^t \times \mathbb{1}_{\text{low}M \,\&\, \text{high}V^q}$
	Newey	Newey	Newey	Newey	Newey	Newey
	Coef. Std. Err. p-value	Coef. Std. Err. p-value	Coef. Std. Err. p-value	Coef. Std. Err. p-value	Coef. Std. Err. p-value	Coef. Std. Err. p-value
User M	0.3697 (0.0231) 0.000***	-0.0899 (0.0262) 0.001***	-0.0436 (0.0175) 0.013**	-0.0611 (0.0171) 0.000***	0.0066 (0.0054) 0.222	-0.0040 (0.0144) 0.782
User V^q	0.0647 (0.0180) 0.000***	0.1597 (0.0294) 0.000***	-0.0647 (0.0171) 0.000***	-0.0688 (0.0133) 0.000***	0.0029 (0.0059) 0.620	0.0099 (0.0140) 0.480
User V^t	0.4220 (0.0518) 0.000***	-0.4861 (0.0805) 0.000***	0.0161 (0.0355) 0.651	-0.2007 (0.0436) 0.000***	-0.0729 (0.0160) 0.000***	-0.1045 (0.0285) 0.000***
User $V^t \times \mathbb{1}_{\text{low}M \,\&\, \text{low}V^q}$	-0.7373 (0.0693) 0.000***	0.1945 (0.0890) 0.029**	0.4812 (0.0973) 0.000***	1.1597 (0.1516) 0.000***	0.0178 (0.0157) 0.257	-0.1331 (0.0513) 0.009***
User $V^t \times \mathbb{1}_{\text{high}M \,\&\, \text{high}V^q}$	-0.1788 (0.0519) 0.001***	0.6639 (0.0772) 0.000***	-0.0078 (0.0454) 0.863	-0.0145 (0.0373) 0.697	0.7867 (0.0579) 0.000***	-0.1284 (0.0472) 0.007***
User $V^t \times \mathbb{1}_{\text{low}M \,\&\, \text{high}V^q}$	-0.7286 (0.0622) 0.000***	1.2800 (0.0955) 0.000***	0.5172 (0.0659) 0.000***	-0.1609 (0.0697) 0.021**	-0.0330 (0.0204) 0.106	1.2642 (0.0809) 0.000***
Pol	0.0362 (0.0145) 0.013**	0.0282 (0.0186) 0.130	0.0220 (0.0173) 0.205	0.0197 (0.0142) 0.164	0.0029 (0.0064) 0.650	-0.0106 (0.0137) 0.438
Fran	-0.0634 (0.0280) 0.023**	0.0432 (0.0248) 0.082*	0.0229 (0.0212) 0.281	0.0269 (0.0160) 0.093*	-0.0099 (0.0102) 0.331	0.0015 (0.0178) 0.934
Firm Fixed Effects	Included	Included	Included	Included	Included	Included
Time Fixed Effects	Included	Included	Included	Included	Included	Included
Num Obs.	15241	15241	15241	15241	15241	15241
Num Panelists	3506	3506	3506	3506	3506	3506
Avg Obs./Panelist	4.3	4.3	4.3	4.3	4.3	4.3
Sanderson-Windmeijer F:	58.43	23.76	26.20	34.35	33.48	23.90

***: significant at 1%, **: significant at 5%, and *: significant at 10% (**two-tailed tests**).

Table OI.2: First-Stage Estimates

	Restaurants (Yearly Panel), Direct Test of Hypotheses											
	DV: M			DV: V^q			DV: V^t			DV: $V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$		
	Newey			Newey			Newey			Newey		
	Coef.	Std. Err.	p-value	Coef.	Std. Err.	p-value	Coef.	Std. Err.	p-value	Coef.	Std. Err.	p-value
User M	0.3774	(0.0232)	0.000***	-0.0918	(0.0265)	0.001***	-0.0506	(0.0175)	0.004***	-0.0643	(0.0183)	0.000***
User V^q	0.0713	(0.0184)	0.000***	0.1828	(0.0304)	0.000***	-0.0697	(0.0179)	0.000***	-0.0601	(0.0180)	0.001***
User V^t	0.3988	(0.0476)	0.000***	-0.5058	(0.0841)	0.000***	0.0364	(0.0387)	0.347	-0.3615	(0.0750)	0.000***
User $V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$	-0.5852	(0.0475)	0.000***	0.6367	(0.0757)	0.000***	0.3607	(0.0504)	0.000***	0.9447	(0.0776)	0.000***
Pol	0.0382	(0.0147)	0.010***	0.0297	(0.0195)	0.127	0.0202	(0.0175)	0.246	0.0106	(0.0177)	0.549
$Fran$	-0.0624	(0.0277)	0.024**	0.0289	(0.0249)	0.245	0.0214	(0.0213)	0.316	0.0173	(0.0232)	0.456
Firm Fixed Effects	Included			Included			Included			Included		
Time Fixed Effects	Included			Included			Included			Included		
Num Obs.	15241			15241			15241			15241		
Num Panelists	3506			3506			3506			3506		
Avg Obs./Panelist	4.3			4.3			4.3			4.3		
Sanderson-Windmeijer F:	109.02			41.52			33.64			40.37		

***: significant at 1%, **: significant at 5%, and *: significant at 10% (**two-tailed tests**).

Table OI.3: First-Stage Estimates

	Hotels (Quarterly Panel), 2x2 Specification					
	DV: M	DV: V^q	DV: V^t	DV: $V^t \times \mathbb{1}_{\text{low}M \,\&\, \text{low}V^q}$	DV: $V^t \times \mathbb{1}_{\text{high}M \,\&\, \text{high}V^q}$	DV: $V^t \times \mathbb{1}_{\text{low}M \,\&\, \text{high}V^q}$
	Newey	Newey	Newey	Newey	Newey	Newey
	Coef. Std. Err. p-value	Coef. Std. Err. p-value	Coef. Std. Err. p-value	Coef. Std. Err. p-value	Coef. Std. Err. p-value	Coef. Std. Err. p-value
User M	0.3885 (0.0195) 0.000***	-0.0389 (0.0173) 0.024**	-0.0996 (0.0206) 0.000***	-0.0870 (0.0183) 0.000***	0.0073 (0.0047) 0.117	-0.0267 (0.0114) 0.019**
User V^q	0.0624 (0.0186) 0.001***	0.2106 (0.0220) 0.000***	-0.1831 (0.0130) 0.000***	-0.0984 (0.0115) 0.000***	0.0046 (0.0052) 0.375	-0.0207 (0.0089) 0.020**
User V^t	0.5692 (0.0357) 0.000***	-0.6253 (0.0356) 0.000***	0.0175 (0.0386) 0.651	-0.2440 (0.0212) 0.000***	-0.1163 (0.0122) 0.000***	-0.2502 (0.0180) 0.000***
User $V^t \times \mathbb{1}_{\text{low}M \,\&\, \text{low}V^q}$	-0.9982 (0.0445) 0.000***	0.3280 (0.0342) 0.000***	0.4630 (0.0612) 0.000***	1.1030 (0.0659) 0.000***	0.0673 (0.0118) 0.000***	0.0215 (0.0207) 0.301
User $V^t \times \mathbb{1}_{\text{high}M \,\&\, \text{high}V^q}$	-0.2176 (0.0355) 0.000***	0.8312 (0.0426) 0.000***	0.0742 (0.0384) 0.053*	-0.0202 (0.0185) 0.274	1.0564 (0.0412) 0.000***	-0.0911 (0.0243) 0.000***
User $V^t \times \mathbb{1}_{\text{low}M \,\&\, \text{high}V^q}$	-0.7700 (0.0378) 0.000***	1.2074 (0.0466) 0.000***	0.6363 (0.0457) 0.000***	-0.0809 (0.0241) 0.001***	0.0273 (0.0131) 0.036**	1.4803 (0.0477) 0.000***
Pol	0.0193 (0.0076) 0.011**	-0.0037 (0.0078) 0.640	-0.0026 (0.0077) 0.740	-0.0042 (0.0063) 0.505	-0.0001 (0.0036) 0.972	0.0040 (0.0060) 0.509
Firm Fixed Effects	Included	Included	Included	Included	Included	Included
Time Fixed Effects	Included	Included	Included	Included	Included	Included
Num Obs.	28393	28393	28393	28393	28393	28393
Num Panelists	1629	1629	1629	1629	1629	1629
Avg Obs./Panelist	17.4	17.4	17.4	17.4	17.4	17.4
Sanderson-Windmeijer F:	43.03	27.24	28.97	49.99	29.03	29.04

***: significant at 1%, **: significant at 5%, and *: significant at 10% (**two-tailed tests**).

Table OI.4: First-Stage Estimates

	Hotels (Quarterly Panel), Direct Test of Hypotheses											
	DV: M			DV: V^q			DV: V^t			DV: $V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$		
	Newey			Newey			Newey			Newey		
	Coef.	Std. Err.	p-value	Coef.	Std. Err.	p-value	Coef.	Std. Err.	p-value	Coef.	Std. Err.	p-value
User M	0.4046	(0.0198)	0.000***	-0.0274	(0.0180)	0.128	-0.1066	(0.0206)	0.000***	-0.1124	(0.0192)	0.000***
User V^q	0.0910	(0.0188)	0.000***	0.2617	(0.0236)	0.000***	-0.1847	(0.0141)	0.000***	-0.1096	(0.0137)	0.000***
User V^t	0.5126	(0.0339)	0.000***	-0.6921	(0.0378)	0.000***	0.0465	(0.0396)	0.241	-0.6038	(0.0400)	0.000***
User $V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$	-0.7569	(0.0332)	0.000***	0.7074	(0.0354)	0.000***	0.4180	(0.0426)	0.000***	1.2035	(0.0452)	0.000***
Pol	0.0199	(0.0078)	0.010***	-0.0025	(0.0082)	0.763	-0.0027	(0.0078)	0.732	-0.0004	(0.0080)	0.957
Firm Fixed Effects	Included			Included			Included			Included		
Time Fixed Effects	Included			Included			Included			Included		
Num Obs.	28393			28393			28393			28393		
Num Panelists	1629			1629			1629			1629		
Avg Obs./Panelist	17.4			17.4			17.4			17.4		
Sanderson-Windmeijer F:	73.03			37.45			34.99			39.60		

***: significant at 1%, **: significant at 5%, and *: significant at 10% (**two-tailed tests**).

OJ Robustness Checks

This appendix reports the two sets of estimation results discussed in the robustness check section in the main paper.

1) Under the alternative assumption that our measures contain missing information, we estimate the coefficients and the standard errors using multiple imputation and the combining rules. This result is reported in Table OJ.1.

2) We assume that all quality and taste deviations are of the same sign, and report the estimation results in Table OJ.2.

Table OJ.1: Estimation Results Using Multiple Imputation

			Restaurants (Yearly Panel)						Hotels (Quarterly Panel)					
			2x2 Specification			Direct Test of Hypotheses			2x2 Specification			Direct Test of Hypotheses		
				Newey			Newey			Newey			Newey	
$log(Rev)$	Prediction:		Coef.	Std. Err.	p-value	Coef.	Std. Err.	p-value	Coef.	Std. Err.	p-value	Coef.	Std. Err.	p-value
M	γ_1		-0.0510	(0.1524)	0.369	-0.0244	(0.1295)	0.425	0.0044	(0.0263)	0.434	0.0139	(0.0227)	0.270
V^q	γ_2	$\gamma_2 < 0$	-0.3835	(0.3971)	0.167	-0.3213	(0.3169)	0.155	-0.1010	(0.0585)	0.042**	-0.0752	(0.0422)	0.037**
V^t	γ_3		-1.3705	(0.6503)	0.018**	-1.3028	(0.5702)	0.011**	-0.1711	(0.1220)	0.080*	-0.1728	(0.1189)	0.073*
$V^t \times \mathbb{1}_{\text{low}M \,\&\, \text{low}V^q}$	γ_4	$\gamma_4 > 0$	0.7571	(0.3975)	0.028**				0.1334	(0.0868)	0.062*			
$V^t \times \mathbb{1}_{\text{high}M \,\&\, \text{high}V^q}$	γ_5	$\gamma_5 > 0$	0.9629	(0.5389)	0.037**				0.1895	(0.1073)	0.039**			
$V^t \times \mathbb{1}_{\text{low}M \,\&\, \text{high}V^q}$	γ_6	$\gamma_6 > 0$	1.0555	(0.6899)	0.063*				0.2121	(0.1243)	0.044**			
$V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$	γ_7	$\gamma_7 > 0$				0.8464	(0.4376)	0.027**				0.1628	(0.0942)	0.042**
Pol			-0.1319	(0.0630)	0.018**	-0.1363	(0.0612)	0.013**	0.0011	(0.0064)	0.432	0.0016	(0.0063)	0.400
$Fran$			0.0028	(0.0803)	0.486	-0.0061	(0.0811)	0.470						
Firm Fixed Effects				Included			Included			Included			Included	
Time Fixed Effects				Included			Included			Included			Included	
Num Obs.				15241			15241			28393			28393	
Num Panelists				3506			3506			1629			1629	
Avg Obs./Panelist				4.3			4.3			17.4			17.4	

***: significant at 1%, **: significant at 5%, and *: significant at 10% (**one-tailed tests**).

Table OJ.2: Robustness Check - Assuming All Deviations are of the Same Sign

			Restaurants (Yearly Panel)								Hotels (Quarterly Panel)							
			2x2 Specification				Direct Test of Hypotheses				2x2 Specification				Direct Test of Hypotheses			
			Newey			*Adjusted*	Newey			*Adjusted*	Newey			*Adjusted*	Newey			*Adjusted*
$log(Rev)$	Prediction:		Coef.	Std. Err.	p-value	p-value	Coef.	Std. Err.	p-value	p-value	Coef.	Std. Err.	p-value	p-value	Coef.	Std. Err.	p-value	p-value
M	γ_1		-0.0769	(0.1466)	0.300	0.225	-0.0607	(0.1263)	0.316	0.316	0.0091	(0.0290)	0.377	0.260	0.0113	(0.0264)	0.334	0.334
V^q	γ_2	$\gamma_2 < 0$	-0.3131	(0.2777)	0.130	0.111	-0.2565	(0.2086)	0.109	0.109	-0.0793	(0.0425)	0.031**	0.039**	-0.0685	(0.0315)	0.015**	0.018**
V^t	γ_3		-1.0856	(0.4437)	0.007***	0.042**	-1.0195	(0.3932)	0.005***	0.015**	-0.1119	(0.0626)	0.037**	0.050*	-0.1028	(0.0594)	0.042**	0.042**
$V^t \times \mathbb{1}_{\text{low}M \,\&\, \text{low}V^q}$	γ_4	$\gamma_4 > 0$	0.5783	(0.2955)	0.025**	0.048**					0.0956	(0.0467)	0.020**	0.043**				
$V^t \times \mathbb{1}_{\text{high}M \,\&\, \text{high}V^q}$	γ_5	$\gamma_5 > 0$	0.5821	(0.3409)	0.044**	0.057*					0.0875	(0.0505)	0.042**	0.052*				
$V^t \times \mathbb{1}_{\text{low}M \,\&\, \text{high}V^q}$	γ_6	$\gamma_6 > 0$	0.7922	(0.4563)	0.041**	0.060*					0.1270	(0.0624)	0.021**	0.043**				
$V^t \times \mathbb{1}_{\text{low}M \text{ or high}V^q}$	γ_7	$\gamma_7 > 0$					0.5917	(0.2927)	0.022**	0.030**					0.0964	(0.0467)	0.019**	0.021**
Pol			-0.1321	(0.0609)	0.015**	0.053*	-0.1373	(0.0596)	0.011**	0.022**	0.0008	(0.0063)	0.451	0.448	0.0010	(0.0063)	0.437	0.437
$Fran$			0.0266	(0.0793)	0.369	0.345	0.0207	(0.0790)	0.397	0.397								
Firm Fixed Effects			Included				Included				Included				Included			
Time Fixed Effects			Included				Included				Included				Included			
Num Obs.			15241				15241				28393				28393			
Num Panelists			3506				3506				1629				1629			
Avg Obs./Panelist			4.3				4.3				17.4				17.4			
First-stage Underidentification Test:			χ^2=36.21 p-val=0.0000				χ^2=62.45 p-val=0.0000				χ^2=68.16 p-val=0.0000				χ^2=87.57 p-val=0.0000			
First-stage Weak Identification Test:			F=20.88				F=44.99				F=53.86				F=109.08			

***: significant at 1%, **: significant at 5%, and *: significant at 10% (**one-tailed tests**).