SESHADRI TIRUNILLAI and GERARD J. TELLIS*

Online chatter, or user-generated content, constitutes an excellent emerging source for marketers to mine meaning at a high temporal frequency. This article posits that this meaning consists of extracting the key latent dimensions of consumer satisfaction with quality and ascertaining the valence, labels, validity, importance, dynamics, and heterogeneity of those dimensions. The authors propose a unified framework for this purpose using unsupervised latent Dirichlet allocation. The sample of user-generated content consists of rich data on product reviews across 15 firms in five markets over four years. The results suggest that a few dimensions with good face validity and external validity are enough to capture quality. Dynamic analysis enables marketers to track dimensions' importance over time and allows for dynamic mapping of competitive brand positions on those dimensions over time. For vertically differentiated markets (e.g., mobile phones, computers), objective dimensions dominate and are similar across markets, heterogeneity is low across dimensions, and stability is high over time. For horizontally differentiated markets (e.g., shoes, toys), subjective dimensions dominate but vary across markets, heterogeneity is high across dimensions, and stability is low over time.

Keywords: consumer satisfaction, quality, dimensions, brand mapping, big data, latent Dirichlet allocation, user-generated content

Online Supplement: http://dx.doi.org/10.1509/jmr.12.0106

Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation

The quality of a product or service is an important determinant of consumer satisfaction, brand performance, and long-term brand success. Prior research has shown that quality drives customer preferences, market share, consumer satisfaction, brand loyalty, price, and, ultimately, firm value (e.g., Jacobson and Aaker 1987; Rust, Zahorik, and Keiningham 1995; Tellis and Johnson 2007; Tellis and Wernerfelt 1987; Tellis, Yin, and Niraj 2009). Managers and researchers usually obtain measures of perceived quality from customers through surveys or interviews, which are typically based on limited samples administered periodically.

With advances in online media and technologies, customers increasingly share their opinions about products on various online platforms such as product reviews, bulletin boards, and social networks (popularly referred to as user-generated content [UGC]). Numerous studies have shown that UGC is influential in determining demand, sales, or financial performance (e.g., Chevalier and Mayzlin 2006; Onishi and Manchanda 2012; Tirunillai and Tellis 2012). Relative to customer surveys, UGC is spontaneous, passionate, widely available, low cost, easily accessible, temporally disaggregate (days, hours, minutes), and live. It is also increasing rapidly and is easier for firms to administer and monitor than surveys. In addition, UGC may be based on

*Seshadri Tirunillai is Assistant Professor, C.T. Bauer College of Business, University of Houston (e-mail: seshadri@bauer.uh.edu). Gerard J. Tellis is Professor of Marketing, Director of the Center for Global Innovation, and Neely Chair in American Enterprise, Marshall School of Business, University of Southern California (e-mail: tellis@usc.edu). Some parts of the study are covered under U.S. Patent 8744896. The study benefited from a grant from Don Murray to the USC Marshall Center for Global Innovation and from the Small Grants Program at University of Houston. Fred Feinberg served as associate editor for this article.

hundreds of thousands of customer contributions on online forums. As such, it tends to represent the "wisdom of the crowds" (Surowiecki 2004). Thus, UGC can serve as a useful source of information or meaning for marketers about consumers' experiences with quality.

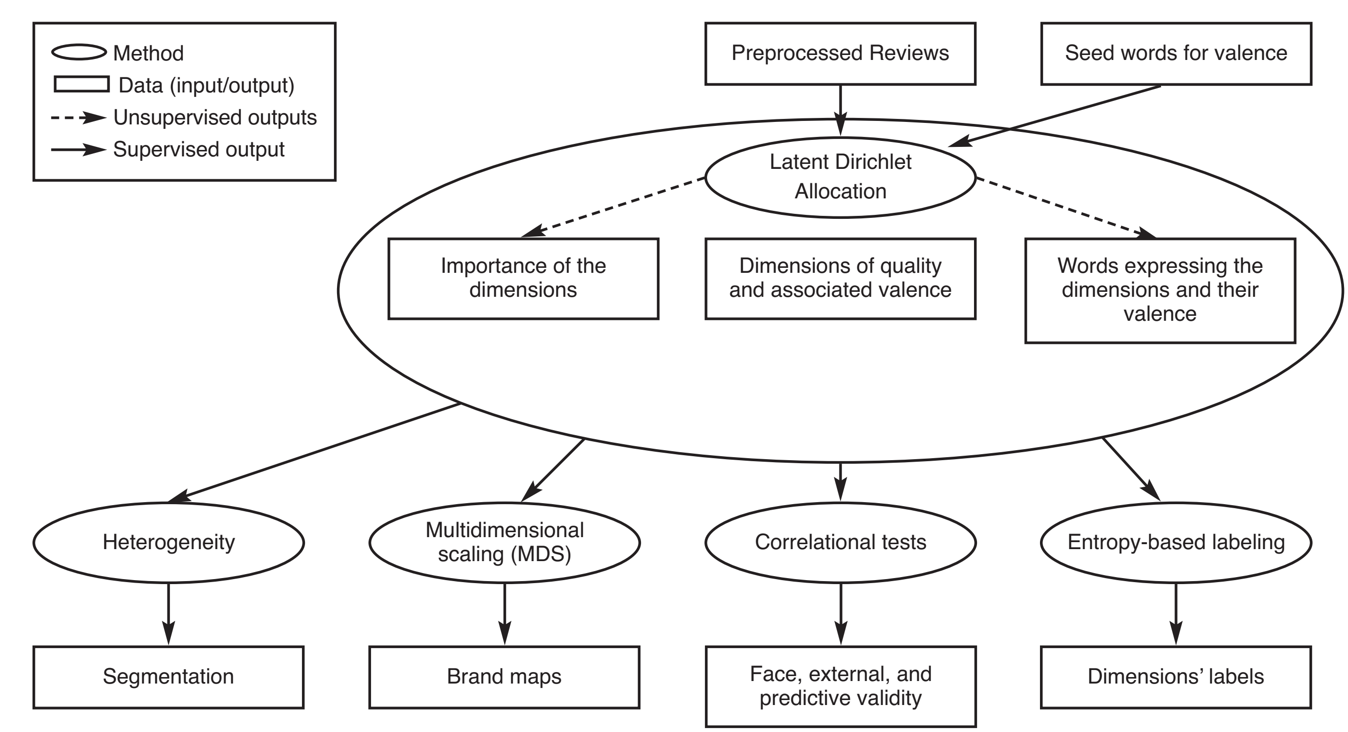
It has long been acknowledged that quality is a multidimensional construct (Klein and Leffler 1981; Mitra and Golder 2006; Tellis and Johnson 2007). The dimensions of quality are critical because they constitute the basis on which consumers evaluate brands and firms compete with one another, design new products, choose brand positioning, and write advertising copy. Traditionally, marketing researchers obtain the latent dimensions of quality through consumer surveys. Latent dimensions (such as performance) are variables that consumers may not explicitly mention but capture or represent a large number of attributes (e.g., the speed, power, or multitasking capabilities of a computer). User-generated content provides a rich source of data to extract the dimensions of quality. This study proposes a unified framework (see Figure 1) for (1) extracting the latent dimensions of quality from UGC; (2) ascertaining the valence, labels, validity, importance, dynamics, and heterogeneity of those dimensions; and (3) using those dimensions for strategy analysis (e.g., brand positioning). Valence is the expression of positive versus negative performance on a dimension or attribute and is termed "sentiment" in text-mining research.

An emerging stream of research in marketing has attempted to explicitly extract or implicitly ascertain the dimensions of quality from UGC. Lee and Bradlow (2011) derive market structure from the product reviews from epinions.com using a constrained optimization approach, exploiting the pros/cons structure of the reviews to extract phrases and assign the words to individual categories. Archak, Ghose, and Ipeirotis (2011) combine multiattribute choice models with multiple text analysis methods to examine the influence of product reviews on consumers' product choice decisions. Netzer et al. (2012) exploit the co-occurrence of words and use semantic network analysis to derive market structure from online consumer forums. Relative to that research stream, the proposed framework in the current study differs in three major ways. First, this study captures the valence expressed in UGC using the unsupervised method, latent Dirichlet allocation (LDA; Blei, Ng, and Jordan 2003), while simultaneously extracting the latent dimensions of quality. Latent Dirichlet allocation–based framework is highly efficient because it can be adapted to handle both big data and highly disaggregate time periods with sparse data. Second, the study extends LDA (typically used for dimension extraction) to extract valence. Third, LDA uses an unsupervised Bayesian learning algorithm to capture context-specific valence (e.g., "small" could be a positive attribute for a mobile phone but a negative attribute when describing a screen). Fourth, the model we use herein does not make assumptions about the structure of the text or the syntactical or grammatical properties of the language, which makes it more suitable to extract latent dimensions in various applications in marketing. The method is also not dependent on assumptions about the underlying distribution of the words, nor is it based on structure of the relationships between the words. Fifth, this study demonstrates the method on a relatively broad sample of five markets and 16 brands, which enables us to make some preliminary generalizations.

The use of LDA over other techniques of text analysis provides the following benefits. First, because LDA effi-

Figure 1

FRAMEWORK FOR UNSUPERVISED PROCESSING OF DIMENSIONS AND VALENCE FOR MARKETING STRATEGY



ciently analyzes data at a highly granular temporal level, it allows for exploration of dynamics over time. In addition, LDA allows for computation of the importance of the extracted dimensions by the intensity of the conversations on each dimension. This function enables the extraction of a parsimonious set of an optimum number of latent dimensions. We can use the results of LDA for further analysis to offer rich managerial insights such as dimensions' importance over time, heterogeneity among consumers' reliance on dimensions, perceptual maps of competing brands, and dynamics of these maps. Latent Dirichlet allocation is one of the base models in the family of "topic models" (Blei 2012) and is flexible enough to undertake such rich analysis. In this study, we use this advantage to extend the LDA model to capture context-specific valence.

We concede that although the core of the LDA model is unsupervised, processing the results to glean managerial insights indeed requires some supervision because their interpretation depends on the market chosen for the analysis. Figure 1 provides a flow chart of the model and illustrates the supervised and unsupervised steps.

Relative to commercial methods for text analysis, our proposed method has the following advantages. It can complete many steps of the analysis using unsupervised methods that involve little human intervention, even labeling dimensions.[1] As a result, it is not necessary for the researcher to know the latent dimensions in advance. Similarly, our method extracts valence from the sample data without requiring client or rater inputs. Most commercial methods require input from clients regarding the positive and negative terms. Thus, the proposed method can process large quantities of data with minimal bias or errors (e.g., tedium) that can occur with human raters.

In summary, this article proposes a unified framework for extracting the latent dimensions of quality, ascertaining the valence on the basis of unsupervised LDA, and determining labels, validity, importance, dynamics, and heterogeneity of those dimensions. The framework also enables us to capture the brand mapping, within-brand segmentation, and examination of the dynamics of brand positions over time.

Specifically, the goal of this study is to answer the following questions:

1. What are the key dimensions of quality expressed in UGC?
2. What is the valence associated with each of these dimensions?
3. What is the validity (face, external, and predictive) of these dimensions?
4. What is the optimum number and importance of these dimensions?
5. How do these dimensions vary across brands in a market and across markets over time?
6. What are the dynamics of these dimensions and the dynamics of brand positions on these dimensions over time?
7. What is the heterogeneity of consumer perceptions and within-brand segments along these dimensions?

The rest of the article is organized as follows. The next four sections describe our method, validation, results, and brand mapping. The final section summarizes the findings, discusses the implications, and lists some limitations of our study.

[1]These methods are popularly referred to as "unsupervised techniques" in the statistics and machine learning literature streams. We use them in this study to analyze large-scale textual data (popularly referred to as "text mining").

METHOD

Sampling

We obtained the data from Tirunillai and Tellis (2012), who collected the data without the help of any market research firm or syndicated data provider. That study used aggregate metrics of the data to predict financial performance. In contrast, the current study takes a deep dive into the content of approximately 350,000 consumer reviews in the data to extract dimensions of quality as well as the valence, validity, importance, optimality, heterogeneity, and dynamics of those dimensions. The Tirunillai and Tellis (2012) study address none of these issues.

The data represent a relatively broad cross-section of categories, which include the following five markets (and brands): personal computing (Hewlett-Packard [HP] and Dell), cellular phones (Motorola, Nokia, Research in Motion Limited [RIM; now BlackBerry Limited], and Palm), footwear (Skechers USA, Timberland Company, and Nike), toys (Mattel, Hasbro, and LeapFrog), and data storage (Seagate Technology, Western Digital Corporation, and SanDisk).

Preparing Text for Statistical Analysis

Analysis of the text in the reviews is difficult for numerous reasons. First, there is no structure in the free-flowing text. Most reviews written by consumers tend to be casual in their word and grammar usage. Second, the textual content in these reviews must be cleansed to remove words that are not informative about the product or its dimensions of quality. Third, many words must be transformed so that they can be manipulated numerically. In this subsection, we summarize the important steps involved in preparing the text for the statistical analysis.

During the preprocessing[2] step, the textual data is cleaned and standardized for analysis.[3] We eliminate non-English characters and words[4] (e.g., HTML tags, URLs, telephone numbers, punctuation) that do not typically have any informational content about the product or the dimensions of quality that we are interested in extracting. We use anaphoric resolution methods to replace the pronouns with the corresponding nouns, especially those of the products or brands. The reviews are broken into individual sentences, usually by the presence of some character signifying the end of the sentence (e.g., ".", "?", "!", the new line character). We apply part-of-speech tagging to retain only words that are adjectives, nouns, or adverbs—that is, words that have information about the product or the product quality. Because these sentences are in a tokenized format (running text converted to individual words or phrases), we replace common negatives of words (e.g., "hardly", "no") by pre-

[2]Theoretically, LDA can be directly applied to the text without preprocessing, but the results have higher error margins and lower reliability and might increase the computational overhead.

[3]We implemented these steps using the modules in Natural Language Toolkit (www.nltk.org)

[4]We eliminated the entire review if more than 80% of the words are not in English. In doing so, we eliminated approximately 3% of the sample of reviews.

fixing a "not" to the token word that follows. After preprocessing, each review is assumed to be an unordered set of words with meaning. We also stem the words (i.e., convert to the root form—e.g., "like" for "likable," "liked," and "liking") using Porter's (1997) stemming algorithm. We remove all stop words (e.g., "the," "and," "when," "is," "at," "which," "on," "in") that are used for connection and grammar but are not required for meaning. We also remove all the words that do not appear in at least 2% of the product reviews in a given market[5] to ensure that the results are not influenced by outlier words rarely used by consumers in expressing opinions about products. The resulting set becomes the "corpus" of text used for further statistical analysis (Manning, Raghavan, and Schütze 2008). We treat each individual review as a separate document and run the aforementioned steps across all the reviews for a given brand in a given market.

Dimension and Valence Extraction

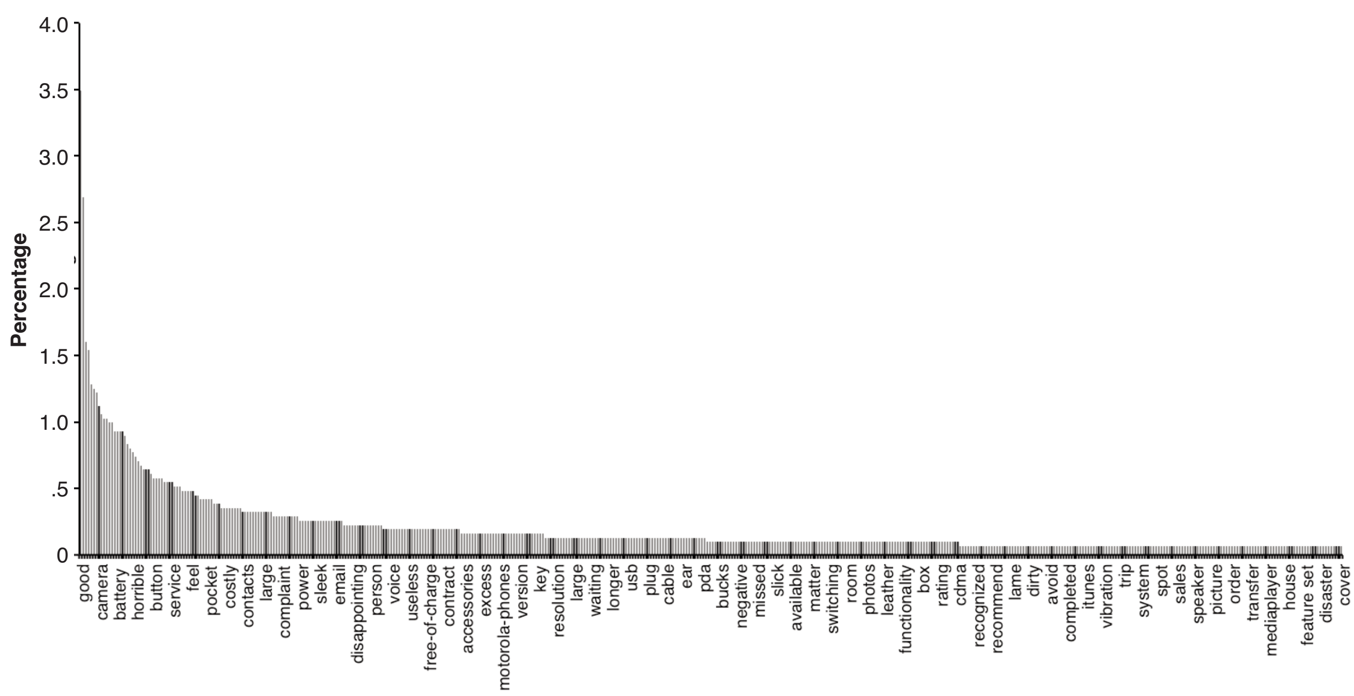
The dimension extraction represents the primary contribution of our study. We explain the dimension extraction in five stages: challenges, intuition, specification, estimation, and labeling.

Challenges. The problem of extracting dimensions of quality from consumer reviews on the Web is analogous to the traditional dimension reduction analysis (e.g., principal component) but presents the following unique challenges. First, a large number of consumers use their own words to describe the quality of the physical and nonphysical attributes. Although some words are common, the corpus of words is very large (numbering in the thousands) and highly skewed, exhibiting characteristics of the long tail (Anderson 2008; see Figure 2), which leads to the "curse of dimensionality." Moreover, consumers express opinions on only those dimensions that are salient to their experience. Thus, each review does not discuss all the dimensions that are salient for all consumers. As a result, the matrix (of reviews × words) in a given time period (week) for a given market used for dimension extraction is very large (e.g., 201 × 2,571, averaged across markets, across the time periods in the sample) yet extremely sparse (containing mostly blanks). Traditional factor-analytic methods do not work reliably on such high-dimensional, sparse matrices because of problems with convergence, parameter stability, and overfitting (Blei, Ng, and Jordan 2003; Buntine and Jakuline 2006). In addition, the valence and adjectives are context specific, are dependent on the product attribute that consumers evaluate, and could reverse for other product attributes even within the same market. For example, a word such as "small" could be evaluated positively in the context of a laptop's size but could have a negative connotation when used in the context of the laptop's memory capacity. Thus, a standard lexicon of positive and negative terms developed across markets may not be applicable for each market.

We exploit recent advances in probability models and Bayesian inference techniques to resolve these two challenges and simultaneously extract the dimensions of quality. We borrow from a class of techniques known as the probabilistic topic models, which are often employed to discover

[5]We also tested the model without these cutoffs. Although there was a steep increase in the computational overhead, the overall results did not change drastically.

Figure 2
WORD DISTRIBUTION (MOBILE PHONE MARKET)



Notes: The vertical axis represents the distribution of each of the words relative to all the words in the sample market.

"topics" (herein used to refer to the dimensions of product quality expressed by consumers) from the textual contents. Specifically, we extend LDA[6] (Blei, Ng, and Jordan 2003) to complete the following steps:

1. Extract valence with dimensions.
2. Identify an optimum number of dimensions.
3. Label the dimensions.
4. Assess the heterogeneity of the dimensions.
5. Position brands on the dimensions.
6. Analyze the dynamics of dimensions and brand positions over time.

Figure 1 portrays the overall framework of our method. First, we use LDA to extract the dimensions, their importance, and the words representing the dimensions from the preprocessed review. Then, we use the output of the LDA for Steps 3–6 (the details of which are outlined in the following sections). To the best of our knowledge, this study is the first use of the method in marketing for dimension extraction from online UGC that incorporates all these issues. Latent Dirichlet allocation is superior to some of the extant methods for extracting dimensions from textual contents for the following reasons. First, it enables joint estimation of valence and dimensions. Second, it is mostly an unsupervised (automated) technique, which implies that researchers do not have to prepare elaborate dictionaries for the analysis. Third, it extracts valence from words on the basis of the context in which the words are used (e.g., the same word can take on different meanings in different markets). Fourth, the latent dimensions are easily interpretable because there is a direct relationship to the attributes (words) that compose the dimensions, allowing for automatic extraction of the candidate words for labeling the dimensions. Fifth, the method is highly efficient and can be extended to handle issues of big data, sparse matrices, and highly disaggregate time periods.

Strictly speaking, the approach can be used to extract both vertically (i.e., objectively) differentiated dimensions (characteristics on which all consumers agree that more is better; e.g., reliability) and horizontally (i.e., subjectively) differentiated dimensions (taste dimensions on which consumers might disagree; e.g., aesthetics). However, because we extract valence with quality, even taste attributes receive a direction or valence in this approach. Thus, even dimensions such as aesthetics appear "aesthetically appealing" in our method.

Intuition. Consumers choose words to express their opinions on one or more dimensions of quality that they believe are worthy of sharing through their review. These dimensions of quality and their associated valence are unobservable (latent) to the researcher. However, each review is a set of words chosen by consumers that can be viewed as representing a random mixture of the latent dimensions. Because we can observe the words, we could infer the latent dimensions from the statistical distribution of these words across all the reviews. Intuitively, words that underlie or describe a dimension will co-occur across the reviews. Thus, observing these co-occurrences and their statistical distribution across the reviews helps us capture the latent dimension and its corresponding valence. Formally, we use LDA for this purpose (Blei 2012; Blei, Ng, and Jordan 2003). The model is a "generative model," which implies that it could be viewed as the intuitive description of the process that generates the review documents on the basis of some probabilistic sampling rules for the hidden parameter (Blei, Ng, and Jordan 2003). Specifically, the model characterizes the process that defines the joint probability distribution over both the observed data (the words in the review) and the hidden random variables (the dimensions of quality). In this sense, the model is an attempt to loosely imitate the process of consumers writing the reviews to retrieve the latent dimensions. Consumers have a finite set of words in their (English-language) vocabulary. While writing the review, they choose the words from their vocabulary to express their opinions on the dimensions of quality. These dimensions have a distribution across the reviews depending on their importance. The model uncovers the distribution of the latent dimensions by beginning with some prior on this distribution. The draws of the words are modeled as a multinomial choice from a finite vocabulary (similar to a consumer choosing the words). We then compute the conditional distribution of the latent variables (dimensions of quality) given the observed variables (words in the review). For statistical inference of the parameters, we try to reverse this generative process and infer the dimensions that most likely generate the observed data. To infer the parameters that best fit the observed sample, we iteratively search over the parameter space of the probability distribution of the words underlying a dimension, the distribution of the dimensions in each of the reviews, and the distribution of the dimensions across all the reviews using Gibbs sampling. Next, we describe the details of the likelihood specification of the generative process and the inference of the dimensions.

Specification. We formally define the dimension of quality to be a latent construct distributed over a vocabulary of words that consumers use to describe their experience with the product.[7] We assume that consumers prefer more to less along these dimensions and attributes (Tellis and Wernerfelt 1987). At the population level (across the corpus of all the reviews), we assume K to be the total number of dimensions that consumers express across all the D reviews [$d \in \{d_1, d_2, \ldots \ldots d_n\}$] of a brand in a given time period. We assume each of the reviews to arise from these latent dimensions (of quality), and each review exhibits a subset of these dimensions in different proportions. A consumer might choose to discuss a subset of these K dimensions in a review by selecting appropriate words that best express his or her experiences with the brand. Because consumers are constrained by the amount of space available to them, and also because of the cost of writing the review (in terms of time and effort), they tend to focus on dimensions of quality that are important in line with their experience of the product. For example, a consumer may devote approximately 40% of the discussion to the ease of use, 30% to the stability, and 15% each to the compatibility and durability dimensions of the product.

[6]Latent Dirichlet allocation is also popular in other disciplines as the "topic model"—that is, the model uncovering the topic of discussion in a given text.

[7]These dimensions are referred to as "topics" in the literature (for an overview of topic models, see Blei 2012).

The kth dimension's importance in the consumer's evaluation of the product may be determined by the proportion allocated to the discussion of that specific dimension that is represented by $\theta_{d,k}$. Following a similar logic for distributional assumptions of valence, we allow $\pi_{d,k}$ to represent the proportion of valence in the dth review. Let $z_{d,n}$ be the dimension and $v_{d,n}$ be the associated valence assigned to $w_{d,n}$, the nth word in the review d, and let N represent all the words in the review. Given our assumption that each of the reviews is a probabilistic mixture of the dimensions of quality and associated valence, the probability of a given word in a given review is

$$(1) \qquad p\left(w_{d,n}\right) = \sum_{k=1}^{K} p\left(w_{d,n} \middle| z_{d,n}, v_{d,n}\right) p\left(z_{d,n}\right) p\left(v_{d,n}\right).$$

In Equation 1, the first term represents the words important in a dimension, and the latter two terms represent the distribution of the dimensions and valence in a given review. The LDA model enables us to combine this equation with a prior probability distribution on θ to provide a complete generative model for the document (Blei, Ng, and Jordan 2003; Griffiths and Steyvers 2004). We briefly describe the likelihood functions of the generative model that can be used to derive the posterior. Let **w**, **z**, and **v** represent the vectors of all words, dimensions, and valence, respectively, across all reviews in the corpus. Let ϕ be the multinomial distribution of the dimension with the associated valence over the vocabulary of the words in the reviews. In addition, we assume α, β, γ to be the hyperparameters on $\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\pi}$ respectively. We assume the draws of words for the dimensions to be from a multinomial distribution of words, and we assume the hyperparameters to be Dirichlet distributed (i.e., the conjugate distribution of these multinomial distributions). Web Appendix A presents details of the draws for the generative model. The following equation represents the joint distribution of the observed words and dimensions (and valence) for the generative process (given the hyperparameters):

$$(2) \qquad p\left(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\pi}, \mathbf{v} \middle| \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\right)$$

$$= \prod_{n=1}^{N} p\left(\mathbf{w_n} \middle| \mathbf{z_n}, \mathbf{v_n}, \boldsymbol{\phi}\right) p\left(\mathbf{z_n} \middle| \boldsymbol{\theta}, \mathbf{v_n}\right) p\left(\mathbf{v_n} \middle| \boldsymbol{\pi}\right) p\left(\boldsymbol{\theta} \middle| \boldsymbol{\alpha}\right) p\left(\boldsymbol{\pi} \middle| \boldsymbol{\gamma}\right) p\left(\boldsymbol{\phi} \middle| \boldsymbol{\beta}\right).$$

In Equation 2, the first three terms include the word-level parameters (in Figure 3), the fourth and fifth terms correspond to review (document-level) parameters, and the last term $[p(\boldsymbol{\phi}|\boldsymbol{\beta})]$ corresponds to the latent dimension (and valence) parameters.[8] Figure 3 shows the graphical representation of the relation between the parameters. We can obtain the likelihood of a review (d_n), which is the probability of jointly observing all the words in a given review, as the marginal distribution of Equation 2.

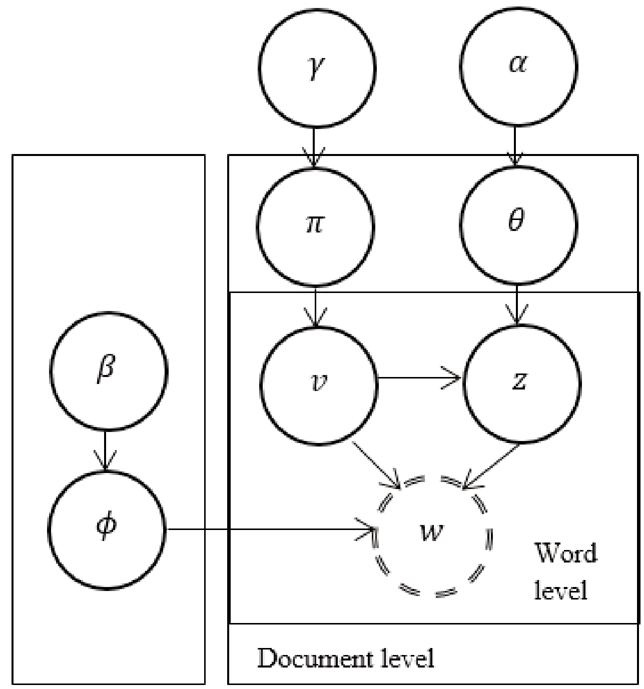
$$(3) \qquad p\left(\mathbf{w} \middle| \alpha, \beta, \gamma\right) = \iiint p\left(\boldsymbol{\phi} \middle| \boldsymbol{\beta}\right) p\left(\boldsymbol{\theta} \middle| \boldsymbol{\alpha}\right) p\left(\boldsymbol{\pi} \middle| \boldsymbol{\gamma}\right)$$

$$\times \prod_{n=1}^{N} p\left(\mathbf{w_n} \middle| \phi_{z,v}\right) p\left(\mathbf{z_n} \middle| \boldsymbol{\theta}\right) p\left(\mathbf{v_n} \middle| \boldsymbol{\pi}\right) d\phi d\theta d\pi.$$

[8]For notational simplicity, we do not include the document-level subscript in the equations.

Figure 3

DIRECTED ACYCLIC GRAPH OF THE VARIABLES AND PRIORS



We could sum across all the z_n and v_n terms and take the product of these marginal distributions to find the likelihood of obtaining p(**w**):

$$(4) \qquad p\left(\mathbf{w} \middle| \alpha, \beta, \gamma\right) = \iiint p\left(\phi \middle| \beta\right) p\left(\theta \middle| \alpha\right) p\left(\pi \middle| \gamma\right)$$

$$\times \prod_{n=1}^{N} p\left(\mathbf{w_n} \middle| \theta, \pi, \phi\right) d\phi d\theta d\pi.$$

We exploit this theoretical generative model for parameter inferences about the latent dimensions in the conditional distribution estimation, outlined subsequently.

Inference and estimation. The main focus in estimation is to infer the distribution of the latent dimensions in a review and the distribution of the words in a dimension. Directly estimating ϕ, the distribution of the words in dimensions, or θ, the distribution of the dimension in each review, can be unreliable (Griffiths and Steyvers 2004). We circumvent this issue by computing the conditional distribution of the latent dimensions given the review, which is the posterior distribution of the assignment of words to the latent dimensions given the words of the review marginalizing ϕ, θ, and π. The posterior distribution for computation of the conditional distribution of the latent dimensions given the observed words in the reviews is given by

$$(5) \qquad \left(\mathbf{z}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\pi}, \mathbf{v} \middle| \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\right) = \frac{p\left(z, w, \phi, \theta, \pi, v, \alpha, \beta, \gamma\right)}{p\left(w \middle| \phi, \theta, \pi, \alpha, \beta, \gamma\right)}.$$

Here, $p(\phi, \theta, \pi, v, z, w, \alpha, \beta, \gamma)$ is the joint probability distribution of all the variables, which can be calculated for any of the latent dimension parameters. The denominator

$p(w|\phi, \theta, \pi, \alpha, \beta, \gamma)$ is the marginal probability distribution, or the probability of observing the review corpus given any parameters of the latent model. In theory, the latter could be calculated by summing across all the possible permutations of assigning the observed words in the corpus to the latent dimensions and valence (as shown in Equation 4). However, in practice, it could be intractable because the number of dimensions could be extremely large (and increasing exponentially with the addition of words in the corpus), as is the case in any Bayesian probabilistic model (Blei, Ng, and Jordan 2003). Thus, we resort to computing the approximation to the posterior (conditional) distribution using Gibb's sampling. The latent variables in the model (θ, π, and ϕ) are estimated by sampling from the posterior conditional distributions of the variables (Griffiths and Steyvers 2004). The posterior is estimated from the Markov chain Monte Carlo procedure adopted for Gibbs sampling.[9] For details, see Web Appendix A.

We identify the valence of the words associated with the dimension in conjunction with the identification of dimension. To implement this, we use an initial small set of seed words that are unambiguously positive (e.g., good, great) or negative (e.g., bad, horrible, lousy) irrespective of the market, context, or dimension (Jo and Oh 2011; Lin and He 2009; Turney and Littman 2003). We use these initial seed words to "train" the model by using a bootstrapping approach. This approach estimates the probability of the valence of the newly encountered words on the basis of the probability of their co-occurrence with the initial seed word. These newly classified words are then appended to the list of the seed words at the start of the next iteration. This process is repeated iteratively until the entire vocabulary of words in the reviews is classified on the basis of the valence.

Selection of optimal number of dimensions. The selection of the optimum number of dimensions for this study is also a model selection problem. We use marginal log-likelihood with fivefold cross-validation to select the optimal number of dimensions. We use the harmonic mean estimator (Newton and Raftery 1994), following prior literature (Griffiths and Steyvers 2004). The number of dimensions the method identifies is decided using the highest posterior likelihood, calculated previously. To determine the optimal number of dimensions for a market, we begin the process by extracting two dimensions and then gradually increase the number of dimensions until the log-likelihood reaches a maximum. For the models with varying dimensions, we sampled the Markov chain Monte Carlo at every hundredth iteration after the log-likelihood value stabilized. Although the harmonic mean estimator is computationally efficient, it is known to suffer from an overestimation problem and low reliability. Therefore, we employed Chib's (1995) estimator following Wallach et al. (2009) to verify the optimal number of dimensions. The numbers of dimensions were not too different in most markets.

Dimension Labeling

Challenges. After estimating the model and extracting the dimensions, we have two tasks. First, we must select the words that better distinguish the reviews associated with that dimension. This ensures that we identify words that occur frequently across the corpus of reviews discussing a specific dimension and sparingly in the reviews that do not discuss the dimension. This criterion is important given that the underlying logic of the method is dependent on words' frequency of occurrence in the reviews.[10] The second task is to assign a label to the given dimension such that it reflects the topic of discussion being evaluated across all the reviews expressing the dimension. These two tasks are interrelated because the words that are important for a given dimension determine its label or provide direction to its labeling.

Intuition. To resolve these two challenges, we derive a score for the word set under a dimension that is based on the mutual information (MI) between the given dimension and the word. Mutual information measures the amount of "information" gained by the given dimension as a result of the presence of the word in that dimension. It reflects the reduction in the amount of uncertainty associated with a dimension due to a given word. A word with high MI has a greater contribution toward that dimension than a word with low MI. Alternatively, we can view MI as a measure of how uniquely a set of words maps to a given dimension. To measure MI, we employ the information theoretic concept of entropy (Grimmer 2010; MacKay 2003). Researchers have used entropy-based measures to assess model fit and to examine the separation between consumer segments (e.g., DeSarbo et al 1992; Kamakura, Kim, and Lee 1996).

Specification. Entropy measures the probability that dimension k generates a randomly chosen review (Manning, Raghavan, and Schütze 2008). Let p be the probability that a randomly chosen review was generated by topic k. We define the entropy that the kth dimension generated a review as

$$(6) \qquad E(k) = -\sum_{\ell=0}^{1} P(\eta = \ell)\log_2 P(\eta = \ell).$$

Here, E(k) refers to the entropy[11] of the given dimension that generated the review, and η represents the event that the review discusses the kth dimension. Intuitively, if all the reviews were generated by a single dimension, E(k) would have the minimum value; if all the dimensions have an equal contribution in generation of the reviews, E(k) would have the maximum value. We can calculate the entropy of the dimension conditional on the words that express the dimension. If a chosen word w* appears in a random review, we model the entropy of the dimension conditional on that word as

$$(7) \quad E(k|w) = -\sum_{\ell=0}^{1}\sum_{w^*=0}^{1} P\left(\eta = \ell \middle| w = w^*\right)\log_2 P\left(\eta = \ell \middle| w = w^*\right).$$

[9]Multiple techniques for computing the approximations to the posterior are available, such as variational inference (Blei, Ng, and Jordan 2003) and collapsed variational inference (Teh et al. 2006). Selection of the method depends on the speed, complexity, and characteristics of the data.

[10]In identifying the dimension labels, we do not use words such as the names of the product or model or descriptions of physical attributes, even if they are extracted within some dimension.

[11]Entropy is typically measured in "bits" and takes nonnegative values.

Estimation. We then calculate how much the word (w*) reduces the uncertainty in the entropy of the dimension using MI. The difference between Equations 6 and 7 expresses the MI gained for dimension k due to word w. If word w reduced the uncertainty of the given dimension, then

$$\text{(8)} \qquad \mathrm{MI}(k|w) = E(k) - E(k|w) \geq 0 \ \forall \ (k, w).$$

If a word provides no information about the topic, the MI is zero. The more information the word contributes to the dimension, the higher its MI score. We then select the top-ranked words such that they cover 90% of the reviews we identified with the given dimension. Thus, we could select the words with higher MI that spanned the dimension. The word(s) that have the highest MI in each dimension could provide possible labels for the given dimension.

VALIDATION

We use multiple validation checks to ascertain the validity of the dimensions of perceived quality. Specifically, we use the following methods to validate the dimensions derived from our (1) face validity with human raters and (2) external validity with *Consumer Reports*.

Face Validity with Human Raters

We compare the results of the automated analysis with those of the dimensions derived by human coders. Two independent trained coders analyzed the reviews for each brand in our sample, similar to the procedure used in prior studies (e.g., Tellis and Johnson 2007). For the content analysis, we developed a set of words that consumers often use to describe the products' dimensions and the associated valence. The coders were given the task of reading each of the reviews to arrive at a set of dimensions and associated valence by parsing the review on the basis of the presence of such terms in each of the reviews. We randomly selected 100 reviews from each brand in a market for this purpose. We compared the coders' decisions with the dimensions derived in the automated analysis to calculate the reliability of the automated analysis. We then used Fleiss' kappa coefficient (Fleiss 1971) to measure the agreement among the two independent human coders and the automated model.

Fleiss' kappa (κ) statistic measures the interrater agreement when there are more than two raters. If the raters assign dimensions on the basis of the topics discussed in the reviews, and p represents the extent to which the raters agree on a given dimension, the κ value is given by

$$\text{(9)} \qquad \kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e}.$$

If p_k represents the proportion of reviews that the raters assigned to a given dimension (k), then $\overline{P}_e$ is computed as $\Sigma_{k=1}^{K} p_k^2$. Furthermore, if P_k measures the extent to which raters agree for the kth dimension across all the reviews d (given by $\{P_k = [1/n(n - 1)]\Sigma_{k=1}^{K} n_{dk}(n_{dk} - 1)\}$), then $\overline{P}$ is computed as the average of all the computed P_k across all the dimensions.

External Validity with Consumer Reports

Consumer Reports is a magazine that evaluates brands on the dimensions deemed important by the expert testers of the products in each market. We restrict our analysis to only markets evaluated by *Consumer Reports*: computers, mobile phones, and footwear.[12] First, we assess the validity of dimensions qualitatively by comparing the dimensions obtained from the automated analysis of UGC with the dimensions evaluated in *Consumer Reports*' ratings of the brands in each of the markets in a given time period. Our aim is to assess the overlap of the dimensions extracted from the automated analysis with that of the dimensions used for rating the brands in the markets in *Consumer Reports*. In the markets for which numerical figures were available, we also assess the correlation between our automated method and *Consumer Reports*' ratings for the dimensions.

RESULTS

In this section, we first summarize the results of the extraction of the dimensions of quality. We then present the validity of these dimensions using multiple methods described previously. Finally, we present the heterogeneity of the dimensions across consumers (reviews) and their stability over time.

Dimensions of Quality

We apply LDA to extract and label the dimensions of quality and the valence of dimensions across all the reviews for each of the 15 firms in our sample. We illustrate the results of the dimensions extracted using snapshots of the brands for the given time period. Table 1 presents the dimensions extracted for Motorola (mobile phone market) during Quarter 4, 2008. It shows that the top six dimensions

[12]Toys are evaluated on a small number of dimensions (e.g., safety aspects) by *Consumer Reports*, so we could not do a comparison.

Table 1
DIMENSIONS OF QUALITY FOR MOTOROLA (MOBILE PHONES, QUARTER 4, 2008)

Instability (Negative)	*Portability (Positive)*	*Receptivity[a] (Positive)*	*Compatibility (Positive)*	*Discomfort[b] (Negative)*	*Secondary Features (Positive)*
Unstable	Smooth	Dependable	Universal	Cramp	Feature
Error	Handy	Reception	Expandable	Big	App
Crash	Portable	Sharp	Supported	Layout	Card
Freeze	Small	Quick	Compatible	Finger	Camera
Reboot	Compact	Crisp	Accessible	Heavy	Wi-Fi

[a]Refers to mobile phone signal.
[b]Refers to discomfort regarding the mobile phone's physical layout.
Notes: The table shows the words with the top MI scores.

are instability, portability, receptivity, compatibility, discomfort, and secondary features.

Table 1 lists the words with the highest MI score relating to each dimension (calculated by Equation 7). These words help label the dimensions and explain the characteristics the dimensions represent. For example, the first column shows the terms relating to the instability dimension. Of these terms, the word with the highest MI score is "unstable." Thus, "instability" could be an appropriate label for this dimension. Following similar logic, the second column characterizes "portability" because it represents words expressing the portable nature of the mobile phone (e.g., "slim," "handy," "portable"). Detailed results of all the markets appear in Web Appendix B (Tables B1–B5).

One of the method's limitations is that for some dimensions, the automatic extraction of the candidate words by MI score for labeling may not convey the words' meaning in its entirety. Each word usually presents partial information of the overall dimension. In such cases, we could resort to manual labeling of the dimensions by human intervention. For each of the dimensions extracted, we randomly select ten reviews that have high posterior probability (from the LDA model) for the dimension. A deeper analysis of the reviews not only helps understand the issue with the specific dimension but also provides more insight into the cause or nature of the associated dimension. For example, all the words in the last column of Table 1 pertain to some secondary features (e.g., applications, additional memory card slot, camera, Wi-Fi) of the Motorola phones that were of significance to consumer experiences; thus, these words co-occur frequently across the reviews.

Some of the dimensions that emerge pertain to issues relating to the retailer and not the manufacturer of the product or the brand discussed. We do not consider these issues dimensions of quality because they do not inherently characterize the product (or the brand).

For words outside the seed words, LDA allocates valence contextually because the model determines the valence depending on the context in which the words colocate. For example, the word "large" has a positive valence in the context of the size of the memory or the screen of a mobile phone; however, the same word has a negative valence in the context of the mobile phone's overall size. This advantage of LDA ensures that we capture words' category-specific valence simultaneously with the analysis of dimensions without needing to develop a category-specific dictionary before doing the analysis. We note the valence associated with the various dimensions in parentheses in Table 1. Positive valence associated with the dimension represents those characteristics of the brand that consumers like, whereas negative valence represents those characteristics that consumers do not like.

To estimate the optimal number of dimensions, we sample the posterior marginal log-likelihood distribution for a varying number of dimensions (as we explain in the "Method" section). For example, Figure 4 shows the plot of the log-likelihood as a function of dimensions for the mobile phone and computer markets. For markets such as mobile phones and computers, the optimal number of dimensions extracted (averaged over the sample period for a given market) are ten and eight, respectively, whereas in markets such as toys and footwear, the average number of dimensions determined are six and eight, respectively.

Figure 4

LOG-MARGINAL LIKELIHOOD FUNCTION (MOBILE PHONE AND COMPUTER MARKETS)

A: Mobile Phones

B: Computers

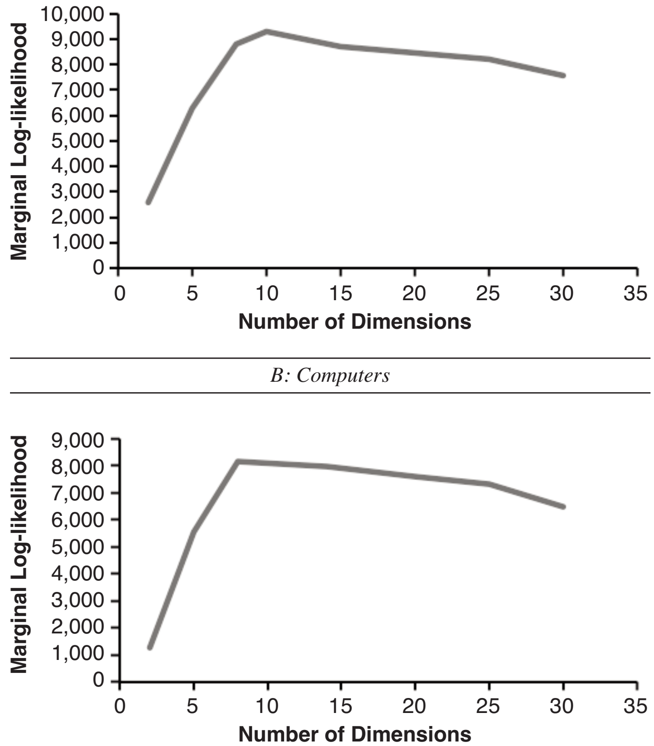


Table 2 compares the top six dimensions across markets and rank-orders the dimensions in these markets. Certain dimensions (e.g., performance, customer service, visual appeal) are common across multiple markets, whereas other dimensions are unique to certain markets (e.g., safety [toys], comfort [footwear], portability [data storage]). These results suggest some preliminary generalizations about the importance of dimensions across categories.[13]

Validation

Face validity with human raters. The average interrater agreement statistic across all the markets in the sample is κ = .59 (the market-specific interrater statistic is 60% [mobile phones], 62% [computers], 57% [data storage], 53% [toys], and 61% [footwear]). Given the nature of the task and the level of ambiguity the raters faced while making decisions, these figures indicate a moderate to substantial agreement between the two raters and the automated analysis (for significance figures for the agreement ratings, see Landis and Koch 1977).

External validity with Consumer Reports. We assess the overlap of the dimensions extracted from the automated analysis with that of the dimensions used by *Consumer*

[13]We check the robustness of these results using a split-sample test. We randomly group reviews from the corpus into two subsamples and run the models on these subsamples. The results (columns 2 and 3 in Table 6) suggest that the method is fairly robust, and the rankings of the split sample are similar to the ranking presented herein.

Table 2
AVERAGE RANKING OF DIMENSIONS ACROSS MARKETS FOR 2005–2009

	Mobile Phones	*Computers*	*Toys*	*Footwear*	*Data Storage*
Performance	3	1	6	6	—
Ease of use	1	2	1	—	—
Visually appealing	4	5	4	2	4
Durability	—	—	5	5	—
Reliability	5	4	—	—	1
Physical support	—	—	3	1	—
Stability	—	3	—	—	2
Portability	—	—	—	—	3
Customer service	6	6	7	4	5
Secondary features	2	7	—	—	6
Safety	—	—	2	—	—
Comfort	—	—	—	3	—

Reports. Extensive studies in economics and marketing have used reports from *Consumer Reports* as dimensions of quality (Mitra and Golder 2006; Tellis and Wernerfelt 1987). Panels A, B, and C of Table 3 show that the dimensions *Consumer Reports* uses overlap considerably with those derived from the automated analysis for the markets in the sample for which *Consumer Reports* had data (mobile phones, computers, and footwear). Note, however, that our automated analysis always included the dimensions of *Consumer Reports*, but the reverse is not true. In particular, in many of the markets, the dimensions reported in *Consumer Reports* are a subset of the dimensions extracted in the automated analysis.[14] This result suggests that crowdsourcing consumer feedback may be more effective and beneficial for both companies and consumers than relying only on experts such as those who help create the reports for *Consumer Reports*. Moreover, expert reviews in sources such as *Consumer Reports* are restricted to a limited number of makes and models within a brand due to the cost and effort involved in evaluating numerous brands. Online product reviews by consumers do not suffer from these restrictions. More importantly, the automated analysis could be obtained weekly or daily, unlike expert reviews of brands, which are undertaken infrequently (usually yearly or once every two years). Thus, managers could use the information in these dimensions closer to real time for their decision making.

We use the Jaccard coefficient to test the degree of overlap between the dimensions evaluated in *Consumer Reports* and those extracted in our automated analysis. If $N(Dim_{lda})$ represents the set of dimensions derived from a set of reviews of a brand from the LDA model and $N(Dim_{CR})$ represents the set of dimensions mentioned in *Consumer Reports* for the same brand in the given time period, we calculate the Jaccard coefficient as

$$(10) \qquad JC = \frac{|N(Dim_{lda} \cap Dim_{CR})|}{|N(Dim_{lda} \cup Dim_{CR})|}.$$

The higher the coefficient's value, the higher the degree of overlap between the two alternate sets of dimensions. The average Jaccard coefficients are .65 (mobile phones), .72 (computers), and .81 (footwear).

[14]In some cases, the dimensions in *Consumer Reports* are more detailed than the automated analysis, and vice versa. In these cases, we take the dimensions that are the closest match between the automated analysis and those of *Consumer Reports*.

Table 3
COMPARISON OF *CONSUMER REPORTS* AND AUTOMATED ANALYSIS

A: Mobile Phones, 2009		
Dimension	*Automated Method*	Consumer Reports
Ease of use (e.g., voice commands, navigation)	✓	✓
Performance (e.g., voice clarity, sensitivity)	✓	✓
Messaging	✓	X
Exhaustibility (battery)	✓	X
Layout discomfort	✓	X
Secondary features (e.g., camera, music player)	✓	✓
Compatibility (e.g., Bluetooth, headphones)	✓	✓
B: Computer, 2008		
Dimension	*Automated Method*	Consumer Reports
Performance	✓	✓
Ease of use	✓	✓
Secondary features (e.g., speaker quality, Wi-Fi)	✓	✓
Compatibility (e.g., Wi-Fi, memory card reader, 64-bit operating system)	✓	✓
Service (e.g., technical support, postsale support, warranty issues)	✓	X
Unreliability (e.g., hard disk failure)	✓	X
Nondurable (e.g., breakage of parts)	✓	X
Portability (size, weight)	✓	X
Ergonomics	✓	✓
C: Footwear, 2006		
Dimension	*Automated Method*	Consumer Reports
Weight	✓	✓
Cushioning	✓	X
Stability	✓	✓
Fit	✓	✓
Flexibility	✓	✓
Breathability	✓	✓
Support	✓	✓
Durability	✓	X
Visual attractiveness	✓	X

Notes: ✓ = included; X = not included.

We also compute the rank-order correlation using the rank of the brands on the dimensions obtained by the automated analysis with the rank on the same dimensions reported by *Consumer Reports* for a given market in a time period. The mean rank-order correlation between the scores is highest for computers (.81), lowest for footwear (.61), and moderate for mobile phones (.74). The differences observed in correlations across categories may be attributed to the varying depth of evaluations by *Consumer Reports* for the different product categories (i.e., the magazine conducted deeper evaluations for mobile phones and computers than for footwear or toys).

Heterogeneity of Dimensions

We assess heterogeneity of dimensions as the dominance of dimensions within a brand in terms of reviews citing them, rather than by the estimated parameters of the dimension. To assess the heterogeneity of consumer perception of these dimensions, we use the Herfindahl index of concentration of reviews mentioning a given dimension within a brand. To do so, we first estimate the percentage share, α, of reviews citing the dimension within a given brand relative to all the other dimensions extracted for the brand. Thus,

$$\alpha = \frac{\text{Total number of reviews citing the dimension}}{\text{Total number of reviews of the brand}}. \quad (11)$$

We then calculate the Herfindahl index[15] of concentration as

$$H = \sum_{i=1}^{n} \alpha^2. \quad (12)$$

The Herfindahl index represents the average concentration of dimensions across all the reviews within a brand. It is an inverse measure of the diversity or heterogeneity in the perception of the dimensions by reviewers.

Column 2 of Table 4 displays the Herfindahl indexes across the brands in the market. Note that the index is relatively high in vertically differentiated markets such as mobile phones, computers, and data storage, with values ranging from 25% to 61%, and relatively low in horizontally differentiated markets such as footwear and toys, with values ranging from 11% to 26%. This finding reflects that, in general, consumers agree more with dimensions in vertically differentiated markets than those in horizontally differentiated markets. This pattern holds even though the absolute number of reviews is high for vertically differentiated markets but low for horizontally differentiated markets.

The reason for this finding is that vertically differentiated markets such as mobile phones and computers have objective dimensions that are relatively well defined for consumers; therefore, consumers' evaluations of products along various dimensions of the brands in these markets are similar and convergent. Thus, a few dimensions reach prominence across all the reviews, revealing little heterogeneity across dimensions (see Table 4, column 2). In contrast, horizontally differentiated markets such as toys or footwear have subjective dimensions on which consumers might have taste differences. Thus, the dimensions exhibit heterogeneity in these markets as reflected in their low scores on the Herfindahl index (see Table 4, column 2). In addition, in horizontally differentiated markets (unlike vertically differentiated markets), the prominent dimensions that contribute

Table 4
HETEROGENEITY OF DIMENSIONS ACROSS BRANDS

Market, Brand	*Herfindahl Index of Concentration*	*Heterogeneity in Dimensions*	*Instability of Herfindahl Index over Time (%)*
Mobile Phones			
Nokia	45.78	Low	3.3
RIM	54.12	Low	3.5
Palm	43.58	Low	2.3
Motorola	48.18	Low	2.1
Computers			
Dell	24.80	Low	1.4
HP	31.68	Low	2.7
Toys			
Hasbro	12.82	Moderate	4.9
Mattel	11.64	High	5.4
LeapFrog	13.58	High	7.6
Footwear			
Timberland	25.74	Moderate	5.1
Skechers	21.52	Moderate	7.4
Nike	23.82	Moderate	8.9
Data Storage			
Seagate	52.44	Moderate	4.8
Western Digital	44.86	Low	3.6
Sandisk	61.02	Low	3.8

to the high Herfindahl index vary across brands within and across markets. These results provide some preliminary generalizations across categories.

Stability of Heterogeneity of Dimensions over Time

The Herfindahl index calculation in the previous section does not take into account the stability or time-varying nature of the heterogeneity of the dimensions. For this purpose, we calculate the percentage instability in Herfindahl index of the dimension[16] over time as follows:

$$V_t = \Delta H_t + 2\left[H_{t-1} - \rho\sigma_t\sigma_{t-1} - \frac{1}{n}\right], \quad (13)$$

where H_t is the Herfindahl index at time t (week) and ρ is the correlation between percentage share of consumers citing the dimension within a given brand relative to all the other dimensions between the two time periods. σ is the standard deviation of shares of dimensions, and n is the total number of dimensions at time t. The overall instability is the average of the weekly instabilities over the four years.

Column 5 of Table 4 shows that the dimensions remain relatively stable over time in vertically differentiated markets such as mobile phones, computers, and data storage, with values ranging from 1% to 4%. However, the dimensions seem relatively more unstable in horizontally differentiated markets such as footwear and toys, with values ranging from 4% to 8%. These results further contribute to generalizations across categories. To test the robustness of the stability of the dimensions over time, we split the sample over the multiple time periods (2005–2007 and 2008–2009) and ran the analysis separately on these subsamples. The results (Table 5) do not suggest a significant change in the dimension over the time.

[15]It is expressed as percentage for our calculation; thus, a Herfindahl index of 10,000 represents one dimension having 100% market share.

[16]Previous research has used similar indexes to assess the mobility in firms' market share over time (e.g., Cable 1997).

Table 5
SPLIT-SAMPLE TEST FOR ROBUSTNESS OF THE STABILITY OF THE DIMENSIONS

Market, Brand	*Instability of Herfindahl Index over Time (%) Sample 2005–2007*	*Instability of Herfindahl Index over Time (%) Sample 2008–2009*
Mobile Phones		
Nokia	3.1	3.5
RIM	3.4	3.7
Palm	2.4	2.6
Motorola	1.8	2.4
Computers		
Dell	1.5	1.8
HP	2.8	2.5
Toys		
Hasbro	5.1	4.7
Mattel	5.2	5.4
LeapFrog	7.8	7.5
Footwear		
Timberland	5.3	5.0
Skechers	7.6	7.3
Nike	8.6	8.5
Data Storage		
Seagate	4.6	5.1
Western Digital	3.8	3.2
Sandisk	3.6	3.9

BRAND MAPPING

This section illustrates an application that uses the results of the LDA model for deriving perceptual maps of the brands in the computer market. Brand mapping consists of graphing the position of competing brands in a market on the basis of their location in space, defined by the key dimensions (e.g., DeSarbo, Grewal, and Scott 2008).

We find the brands' positions on a brand space by mapping the "distance" between the brands on a given dimension. The distance between two brands is estimated as the distributional similarity or dissimilarity between the vector of words underlying the dimensions. More specifically, we measure the distance between brands a and b as $\text{distance}_{a,b} = f(\theta_k^a, \theta_k^b)$, where the function f() calculates the Hellinger distance (Rao 1995) between the probability distributions of the words with the highest MI underlying the dimensions ($\boldsymbol{\theta}_k$) across all the reviews of the two brands (a, b). We use the Hellinger distance[17] measure because research has proven it to be superior to some of the traditional techniques (e.g., correspondence analysis) for measuring the distance between probability distributions (Cuadras, Cuadras, and Greenacre 2006; Rao 1995). Hellinger distance for continuous probability measures A and B is defined as

$$f\left(\theta_k^a, \theta_k^b\right) = \left[\frac{1}{2}\int\left(\sqrt{\frac{dA}{dx}} - \sqrt{\frac{dB}{dx}}\right)^2 dx\right]^{\frac{1}{2}}. \tag{14}$$

In the case of discrete distributions, the Hellinger distance between the probability distribution of two brands on any given dimension is calculated as

$$f\left(\theta_k^a, \theta_k^b\right) = \left[\frac{1}{2}\sum_k\left(\sqrt{\theta_k^a} - \sqrt{\theta_k^b}\right)^2\right]^{\frac{1}{2}}. \tag{15}$$

This is related to the Euclidian norm of the difference in the square root vectors of the discrete probability distributions. We measure the Hellinger distance for all combinations of brands within a given market to derive the similarity matrix of the brands. For example, if there are three brands, we get the similarity matrix between the three brands by calculating the six $\left[C\binom{3}{2}\right]$ combinations. Using the similarity matrices of brands derived from Equation 15, we can map the brands' positions using commonly available multivariate techniques such as multidimensional scaling (MDS) (Desarbo, Ramaswamy, and Lenk 1993; DeSarbo, Young, and Rangaswamy 1997; Rao 1995). Next, we describe four aspects of mapping: static brand mapping, within-brand segmentation, dynamic brand segmentation, and dynamics of dimensions.

Static Brand Mapping

The static brand map consists of carrying out the aforementioned procedure for brands on the top dimensions for a fixed period of time. Figure 5, Panels A, B, and C, presents perceptual maps for three markets (mobile phones, computers, and toys) during Quarter 4 (October to December) of 2008. For the purpose of illustration, we chose the top two most-important dimensions on the basis of the frequency of occurrence of these dimensions across all the reviews in the given time period for these markets. Figure 5, Panel A, shows that the brands Motorola and BlackBerry (under RIM) are rated better than Palm and Nokia on the dimension of performance. For computers (Figure 5, Panel B), HP does better than Dell on both performance and ease of use dimensions. For toys (Figure 5, Panel C), Mattel outperforms LeapFrog and Hasbro in terms of safety and durability dimensions.

Within-Brand Segmentation

We adopt a vector-dimensional procedure to examine the within-brand segments[18] (e.g., DeSarbo, Ramaswamy, and Lenk 1993; Wedel and DeSarbo 1996). In this segmentation approach, more on each dimension is preferable to less (in contrast to the ideal-point segmentation approach, in which segments have an ideal point and the brands closer to the ideal point are preferred by the segment). The vector-based approach to segmentation assumes that the ideal point (or segment) is at an infinite distance in dimensional space.

In our data, consumers describe in detail the characteristics that are important to them, which lead them to allocate more words to these dimensions. We segment consumers on the basis of the proportion of words they allocate to the vari-

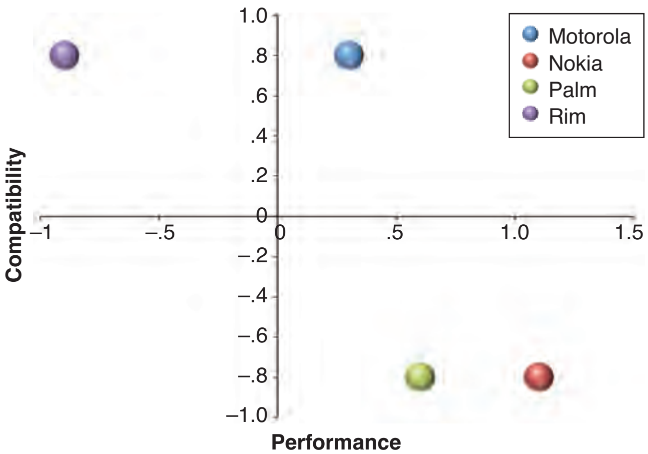
[17]We do not use divergence measures such as the Kullback–Liebler statistic (Lee 1999), because we are interested in deriving a metric distance measure that can be easily used to calculate the perceptual maps. The Hellinger metric, unlike the Kullback–Liebler divergence metric, is well suited for this purpose because it is a symmetric and nonnegative measure and thus can be interpreted as a distance measure (Rao 1995).

[18]Conventional segmentation approaches use data in which the consumers evaluate multiple brands. In the current research, we base our approach on consumer evaluation of a single brand and infer the segments and heterogeneity within the brand. We use the term "segmentation" to refer to the consumers' within-brand segmentation.

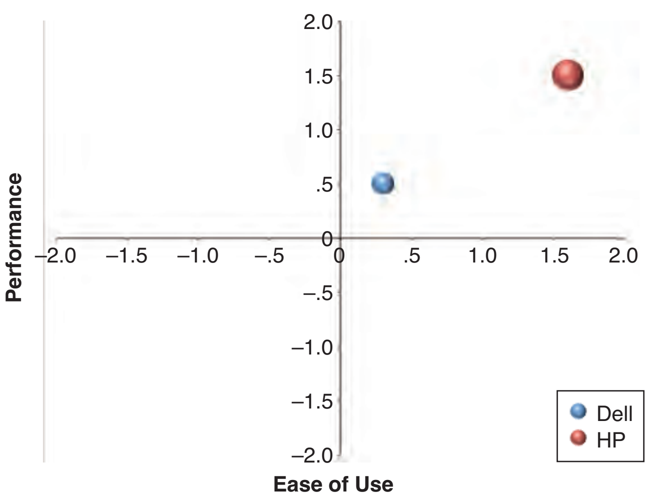
Figure 5

BRAND POSITIONING ON TWO DIMENSIONS OF QUALITY

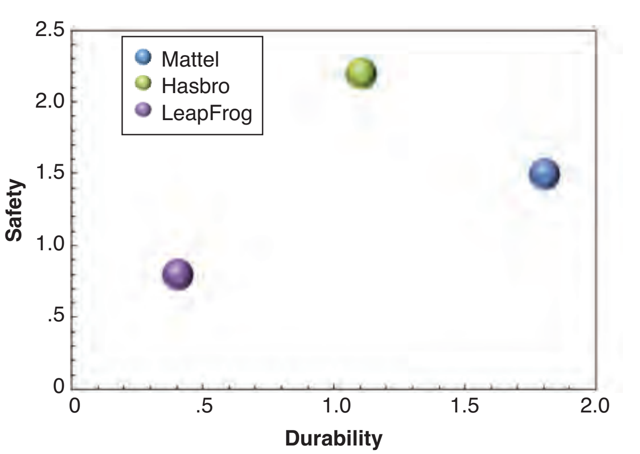
A: Mobile Phone Brands



B: Computers Brands



C: Toy Brands



ous dimensions of quality in their reviews. We calculate the associated segment size by the volume of reviews citing these dimensions. Figure 6 shows the size and location of segments on two of these dimensions, ease of use and performance, in the computer and footwear markets.

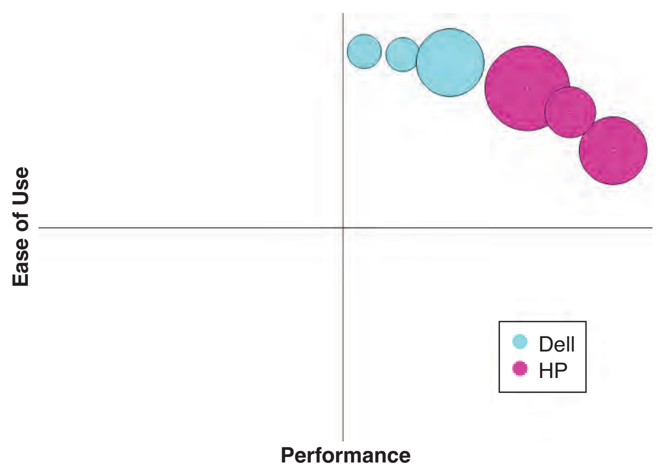
Dynamic Brand Mapping

This section describes the dynamics of brands on dimensions and the dynamics of the dimensions themselves. To capture the dynamics of the dimensions over any time period, we run the analysis on each week during that period (e.g., between June 2005 and December 2010). We use the dimensions extracted from the analysis of the entire pool of reviews (and the words extracted for each of these dimensions) as priors to run this analysis in each of the time periods. Specifically, we extract the probability mass of each of the dimensions occurring in a given week. We define the estimated probability that a dimension k occurred in the review d in time period t as

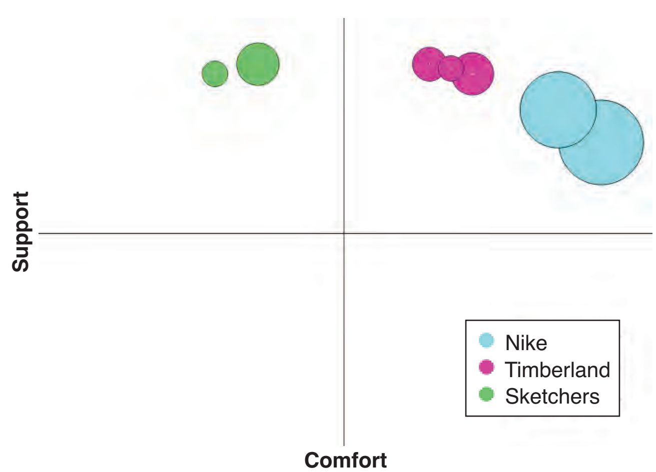
Figure 6

WITHIN-BRAND SEGMENTATION OF MARKETS USING THE DIMENSIONS EXTRACTED

A: Computer Market, 2005–2006



B: Footwear Market, 2005–2006



$$(16) \qquad \hat{p}(k|t=\tau) = \sum_{d|t_d=\tau} \hat{p}(k|d)\hat{p}(d|t=\tau).$$

We illustrate this in the computer market. Figure 7 traces the evolution of the brand position of Dell (depicted in blue) and HP (depicted in green) during the sample period (June 2005 through December 2009) on a weekly basis. Because the dynamics of the dimensions cannot be easily depicted on paper, we also highlight the changes over time using motion charts (see the Web Appendix). The two axes correspond to the scaled probability mass of the brands measured for the dimensions of ease of use and performance. These dimensions emerge as the most frequently discussed dimensions in these markets during the time period. Both these brands are evaluated along these dimensions over the time period, and the positions are depicted in the latent space corresponding to these two dimensions.

As the chart illustrates, Dell's position on ease of use is more unstable and changes rapidly over the time period, indicating that consumer opinion on Dell's ease of use dimension is relatively volatile. In contrast, HP's evolution is more stable along these two dimensions in the same time period, in line with our expectations. During early to mid-2005, Dell was prominent in the news for bad product performance and customer service (e.g., Jeff Jarvis's popular blog about Dell's poor customer service and product quality[19]). Dell's subsequent response was to open the Dell Direct online forums to improve its customer interface, service, and initiatives to improve product quality. The increase in positive opinion about Dell in online reviews regarding its ease of use and performance is visible in the chart as of Week 3, 2007.

[19]See http://buzzmachine.com/archives/cat_dell.html.

The dimensions that are salient for the customers do not undergo drastic change over time, a result we can infer from the split-sample test. Following prior studies (e.g., Lee and Bradlow 2011), we divided the sample into two time period samples: from 2005 to 2007 and 2007 to 2009. We then ran the analysis on these two samples. The results for the mobile phone market appear in Table 6 (columns 5 and 6). The ease of use and portability dimensions were absent in the 2005–2007 period but became important to customers in 2008–2009. Similarly, visual appeal was of immense importance in 2005–2007 but was not much favored in the 2008–2009 period. Notably, the dimension of efficiency (of power) has been captured as important throughout our analysis, and *Consumer Reports* included it for mobile phone ratings in 2010. Similarly, the importance of the secondary features dimension increased over time, whereas "receptivity" decreased over the time period in our analysis. This is also reflected in *Consumer Reports*' inclusion of features such as display size, voice command, and navigation and exclusion of dimensions such as "sensitivity" in 2010, occurrences that are in line with our results. For other dimensions, the order of the ranking does not vary much over the two time periods across all the dimensions, suggesting that the customer-perceived dimensions for brands are relatively strong in this market. Similar results can be observed in other markets (for the results of the computer and footwear markets, see Web Appendix C).

Figure 7

EVOLUTION OF THE POSITION OF THE BRANDS ON THE EASE OF USE AND PERFORMANCE DIMENSIONS IN THE COMPUTER MARKET DURING 2005–2010 (WEEKLY)

Table 6
SPLIT-SAMPLE TEST FOR ROBUSTNESS OF THE DIMENSIONS

	Ranking of the Top Dimensions Among the Samples (Mobile Phone Market)				
	Sample 1	*Sample 2*	*Entire Sample*	*2005–2007 Sample*	*2008–2009 Sample*
Ease of use					*2*
Efficiency (power)	*5*	4	4	4	4
Comfort	*4*	5	5	5	
Stability	3	3	3	3	*5*
Portability	1	1	1		1
(Secondary) features	6	6	6	6	*3*
Visual appeal				1	
Receptivity	2	2	2	2	*6*

Notes: Samples 1 and 2 refer to the split-sample study, in which reviews were sampled randomly from the entire corpus. The 2005–2007 sample and the 2008–2009 sample refer to the sample split to assess the stability of the dimensions over time. Rankings in italics denote change.

Dynamics of Dimensions

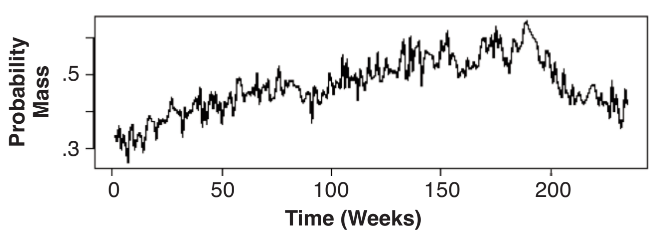
For the dimensions of quality that vary over time, we can assess the continuous evolution to obtain more fine-grained insights. In most markets, the brand trajectories on the dimensions of quality seem to evolve smoothly over a period of time; however, in some markets there is turbulence in certain time periods, as reflected by the transient spurts and falls of brands along these dimensions.[20] This can be partially attributed to exogenous shocks due to product launches in the same category, launch failures of some models of the product, or the introduction of competing products.

We illustrate the dynamics of the dimensions using the mobile phone market. Figure 8 shows the evolution of the ease of use dimension for mobile phones. To aid visualization, in Panel A, we present the probability mass of the dimension as estimated for the mobile phone market (specifically, BlackBerry) from Equation 16 and present the volume of reviews mentioning the given dimension in Panel B. These values increase over the time period to mid-2008 and then gradually decline. Some of the trends and transient spikes can be associated with the launch of new products. For example, the increasing trend from week 150 can be attributed to the release of the BlackBerry Storm smartphone (released in December 2008). The opposite effect occurs when strong competitive products are introduced. This effect is evident in the decline of the BlackBerry brand on the ease of use dimension at approximately week 180. This result can partially be attributed to the increasing penetration of iPhone in the smartphone market, which could have caused increased expectations of BlackBerry phones' ease of use. In this case, the trajectory of a brand's dimensions of quality seems to be related to the entry and exit of other brands in the market. These trends may be attributed to innovation in the underlying technologies. In the smartphone market, radical technological advances in these areas (e.g., touch screens, voice recognition) influence consumer perceptions of the product. Notably, these changes are also associated with an increase in the volume of discussion of these dimensions around the same time period (as illustrated in Figure 8, Panel B). We find similar results in other vertically differentiated markets. The horizontally differentiated markets are fairly immune to these volatilities and exhibit more gradual changes over time.

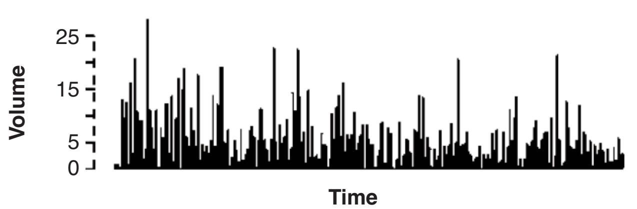
[20]These spikes in brand positions preclude us from employing techniques that would enable us to embed time-varying parameters in the model (e.g., the dynamic topic model; Blei and Lafferty 2006) because they penalize large transient changes occurring over time.

Figure 8
VARIATION IN THE EASE OF USE DIMENSION FOR THE MOBILE PHONE MARKET (BLACKBERRY)

A: Probability Mass Associated with the Ease of Use Dimension



B: Frequency of the Ease of Use Dimension Across Reviews in a Given Week



DISCUSSION

Summary of Findings

The main findings of the study are as follows:

1. Online chatter (i.e., UGC) is rich in marketing meaning. This meaning can be distilled by extracting key latent dimensions of consumer satisfaction about the quality of brands. For this purpose, a few dimensions are adequate.
2. Dimensions differ across various brands in a given market and across markets. Some dimensions (e.g., ease of use, performance, visual appeal) are important across multiple markets, whereas other dimensions (e.g., safety, receptivity, physical support) are important only in certain markets.
3. The valence associated with the dimensions varies across markets.

4. These dimensions exhibit face validity with respect to dimensions extracted by independent human raters as well as external validity with respect to dimensions listed in *Consumer Reports*.
5. Multidimensional scaling can capture brands' positions on dimensions. The brands' positions along these dimensions change over time, and dimensions themselves change in importance over time.
6. For vertically differentiated markets (e.g., mobile phones, computers), objective dimensions rank high, heterogeneity is low across dimensions, and stability is high over time. For horizontally differentiated markets, subjective dimensions rank high, heterogeneity is high across dimensions, and stability is low over time.

Contributions

This study proposes a unified framework to extract latent dimensions from rich user-generated data. It makes several contributions to the literature. First, the framework captures the valence expressed in UGC while simultaneously extracting the latent dimensions of quality using partly automated methods. Second, the framework efficiently analyzes the dynamics of experienced quality at a highly granular temporal level. Third, the framework shows the importance of the extracted dimensions by the time-varying intensity of the conversations on each dimension. Fourth, the framework extracts a parsimonious set of an optimum number of latent dimensions of quality; thus, the number of dimensions does not need to be fixed or known a priori. Fifth, the framework estimates the heterogeneity among consumers on the dimensions. Sixth, this framework demonstrates the method on a relatively broad sample of five markets and 16 brands, enabling us to make some preliminary generalizations.

Implications

This study has many valuable implications for managerial practice. First, it enables managers to ascertain the valence, labels, validity, importance, dynamics, and heterogeneity of latent dimensions of quality from user-generated data. Second, it enables managers to observe how brands compete on multidimensional space. Third, it enables managers to track how this competition varies over time in great detail. This function is currently available at the weekly level, but in the future, it will be available at the daily level. Finally, the dimensions of quality can be a basis for determining consumer satisfaction, brand ranking, new product design, and ad content design.

Limitations and Further Research

This study has some important limitations. First, the models used to extract the latent dimensions of quality are computationally intensive. However, with the current advances in computing and the increasing adoption of large-scale computing techniques, this limitation will dissipate over time. Moreover, adoption of alternate estimation methods such as variational Bayesian inference could also help reduce the time complexity. Second, this study focuses only on product reviews, but it could be extended to other forms of textual communication (e.g., online forums of products, blogs and microblogs, mobile conversations such as tweets). For this purpose, researchers can use preprocessing of text by procedures outlined previously with minor modifications as per the platform. It could be further extended to extract latent topics from news reports, financial documents, advertisement copy, and other textual documents that marketing scholars often use. Third, the LDA model is sensitive to the values of the hyperparameter of the Bayesian priors, which could influence the results in terms of the number of dimensions extracted. Fourth, we neither include marketing mix variables (e.g., advertising, promotions) nor study their impact on the brands or dimensions. Fifth, we do not analyze rare or infrequent words in the long tail of the distribution; these words could reflect emerging consumer preferences that could be very helpful in new product design. Sixth, the model's parameter space could be extended to include other variables in the parameters, such as time, product, or consumer characteristics that can help account for temporal dependencies, product differentiation, and heterogeneity, respectively. These limitations could be rich avenues for further research.

REFERENCES

Anderson, Chris (2008), *The Long Tail: Why the Future of Business Is Selling Less of More*. New York: Hyperion Books.

Archak, Nikolay, Anindya Ghose, and Panagiotis Ipeirotis (2011), "Deriving the Pricing Power of Product Features by Mining Consumer Reviews," *Management Science*, 57 (8), 1485–1509.

Blei, David M. (2012), "Probabilistic Topic Models," *Communications of the ACM*, 55 (4), 77–84.

——— and John D. Lafferty (2006), "Dynamic Topic Models," in *Proceedings of the 23rd International Conference on Machine Learning*. New York: Association for Computing Machinery, 113–20.

———, Andrew Y. Ng, and Michael I. Jordan (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022.

Buntine, Wray and Aleks Jakulin (2006), "Discrete Component Analysis," in *Subspace, Latent Structure and Feature Selection*, C. Saunders, M. Grobelnik, S. Gunn, S., and J. Shawe-Taylor, eds. New York: Springer, 1–33.

Cable, John R. (1997), "Market Share Behavior and Mobility: An Analysis and Time-Series Application," *Review of Economics and Statistics*, 79 (1), 136–41.

Chevalier, Judith and Dina A. Mayzlin (2006), "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research*, 43 (August), 345–54.

Chib, Siddhartha (1995), "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90 (432), 1313–21.

Cuadras, C., Cuadras, D., and Greenacre, M. (2006), "A Comparison of Different Methods for Representing Categorical Data," *Communications in Statistics—Simulation and Computation*, 35 (2), 447–59.

DeSarbo, Wayne S., Rajdeep Grewal, and Crystal J. Scott (2008), "A Clusterwise Bilinear Multidimensional Scaling Methodology for Simultaneous Segmentation and Positioning Analyses," *Journal of Marketing Research*, 45 (June), 280–92.

———, Venkatram Ramaswamy, and Peter Lenk (1993), "A Latent Class Procedure for the Structural Analysis of Two-Way Compositional Data," *Journal of Classification*, 10 (2), 159–93.

———, Michel Wedel, Marco Vriens, and Venkatram Ramaswamy (1992), "Latent Class Metric Conjoint Analysis," *Marketing Letters*, 3 (3), 273–88.

———, Martin R. Young, and Arvind Rangaswamy (1997), "A Parametric Multidimensional Unfolding Procedure for Incomplete Nonmetric Preference/Choice Set Data in Marketing Research," *Journal of Marketing Research*, 34 (November), 499–516.

Fleiss, J.L. (1971), "Measuring Nominal Scale Agreement Among Many Raters," *Psychological Bulletin*, 76 (5), 378–82.

Griffiths, Thomas L. and Mark Steyvers (2004), “Finding Scientific Topics,” *Proceedings of the National Academy of Sciences of the United States of America*, 101 (Supplement 1), 5228–35.

Grimmer, Justin (2010), “A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases,” *Political Analysis*, 18 (1), 1–35.

Jacobson, Robert and David A. Aaker (1987), “The Strategic Role of Product Quality,” *Journal of Marketing*, 51 (October), 31–44.

Jo, Yohan and Alice Oh (2011), “Aspect and Sentiment Unification Model for Online Review Analysis, in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. New York: Association for Computing Machinery, 815–24.

Kamakura, Wagner A., Byung-Do Kim, and Jonathan Lee (1996), “Modeling Preference and Structural Heterogeneity in Consumer Choice,” *Marketing Science*, 15 (2), 152–72.

Klein, Benjamin and Keith B. Leffler (1981), “The Role of Market Forces in Assuring Contractual Performance,” *Journal of Political Economy*, 89 (4), 615–41.

Landis, J. Richard and G.G. Koch (1977), “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, 33 (1), 159–74.

Lee, Lillian (1999), “Measures of Distributional Similarity,” in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. New York: Association for Computing Machinery, 25–32.

Lee, Thomas and Eric Bradlow (2011), “Automated Marketing Research Using Online Customer Reviews,” *Journal of Marketing Research*, 48 (October), 881–94.

Lin, Chenghua and Yulan He (2009), “Joint Sentiment/Topic Model for Sentiment Analysis,” in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*. New York: Association for Computing Machinery, 375–84.

MacKay, David (2003), *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press.

Manning, Christopher D., Prabhaker Raghavan, and Hinrich Schütze (2008), *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press.

Mitra, Debanjan and Peter N. Golder (2006), “How Does Objective Quality Affect Perceived Quality? Short-Term Effects, Long-Term Effects, and Asymmetries,” *Marketing Science*, 25 (3), 230–47.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko (2012), Mine Your Own Business: Market-Structure Surveillance Through Text Mining,” *Marketing Science*, 31 (3), 521–43.

Newton, Michael A. and Adrian E. Raftery (1994), “Approximate Bayesian Inference with the Weighted Likelihood Bootstrap,” *Journal of Royal Statistical Society (Series B)*, 56 (1), 3–48.

Onishi, Hiroshi and Puneet Manchanda (2012), “Marketing Activity, Blogging and Sales,” *International Journal of Research in Marketing*, 29 (3), 221–34.

Porter, M.F. (1997), “An Algorithm for Suffix Stripping,” in *Readings in Information Retrieval*, Karen Sparck Jones and Peter Willett, eds. San Francisco: Morgan Kaufmann Publishers, 313–16.

Rao, C. Radhakrishna (1995), “Use of Hellinger Distance in Graphical Displays,” in *Multivariate Statistics and Matrices in Statistics*. Leiden, The Netherlands: Brill Academic Publisher, 143–61.

Rust, Roland T., Anthony J. Zahorik, and Timothy L. Keiningham (1995), “Return on Quality (ROQ): Making Service Quality Financially Accountable,” *Journal of Marketing*, 59 (April), 58–70.

Surowiecki, James (2004), *The Wisdom of Crowds*. New York: Random House.

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, and David M. Blei (2006), “Hierarchical Dirichlet Processes,” *Journal of the American Statistical Association*, 101 (476), 1566–81.

Tellis, Gerard J. and Joseph Johnson (2007), “The Value of Quality,” *Marketing Science*, 26 (6), 758–73.

——— and Birger Wernerfelt (1987), “Competitive Price and Quality Under Asymmetric Information,” *Marketing Science*, 6 (3), 240–53.

———, Yiding Yin, and Rakesh Niraj (2009), “Does Quality Win? Network Effects Versus Quality in High-Tech Markets,” *Journal of Marketing Research*, 46 (April), 135–49.

Tirunillai, Seshadri and Gerard J. Tellis (2012), “Does Chatter Really Matter? Dynamics of User-Generated Content and Stock Performance,” *Marketing Science*, 31 (2), 198–215.

Turney, Peter D. and Michael L. Littman (2003), “Measuring Praise and Criticism: Inference of Semantic Orientation from Association,” *ACM Transactions on Information Systems (TOIS)*, 21 (4), 315–46.

Wallach, Hanna M., Iaian Murray, Ruslan Salakhutdinov, and David Mimno (2009), “Evaluation Methods for Topic Models,” in *Proceedings of the 26th Annual International Conference on Machine Learning*. New York: Association for Computing Machinery, 1105–1112.

Wedel, Michel and Wayne S. DeSarbo (1996), “An Exponential-Family Multidimensional Scaling Mixture Methodology,” *Journal of Business & Economic Statistics*, 14 (4), 447–59.