Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings

BART DE LANGHE
PHILIP M. FERNBACH
DONALD R. LICHTENSTEIN

This research documents a substantial disconnect between the objective quality information that online user ratings actually convey and the extent to which consumers trust them as indicators of objective quality. Analyses of a data set covering 1272 products across 120 vertically differentiated product categories reveal that average user ratings (1) lack convergence with *Consumer Reports* scores, the most commonly used measure of objective quality in the consumer behavior literature, (2) are often based on insufficient sample sizes which limits their informativeness, (3) do not predict resale prices in the used-product marketplace, and (4) are higher for more expensive products and premium brands, controlling for *Consumer Reports* scores. However, when forming quality inferences and purchase intentions, consumers heavily weight the average rating compared to other cues for quality like price and the number of ratings. They also fail to moderate their reliance on the average user rating as a function of sample size sufficiency. Consumers' trust in the average user rating as a cue for objective quality appears to be based on an "illusion of validity."

Keywords: online user ratings, quality inferences, consumer learning, brand image, price-quality heuristic

Consumers frequently need to make a prediction about a product's quality before buying. These predictions are central to marketing because they drive initial sales, customer satisfaction, repeat sales, and ultimately profit, as well as shareholder value (Aaker and Jacobson 1994; Bolton and Drew 1991; Rust, Zahorik, and Keiningham 1995). Before the rise of the Internet, consumers' quality predictions were heavily influenced by marketer-controlled variables such as price, advertising messages, and brand name (Erdem, Keane, and Sun 2008; Rao and Monroe 1989). But the consumer information environment has changed radically over the last several years. Almost all retailers now provide user-generated ratings and narrative reviews on their websites, and the average user rating has become a highly significant driver of sales across many product categories and industries (Chevalier and Mayzlin 2006; Chintagunta, Gopinath, and Venkataraman 2010; Loechner 2013; Luca 2011; Moe and Trusov 2011; for a recent meta-analysis, see Floyd et al. 2014).

Most people consider the proliferation of user ratings to be a positive development for consumer welfare. User

All authors contributed equally and are listed in alphabetical order. Bart de Langhe (bart.delanghe@colorado.edu) and Philip M. Fernbach (philip.fernbach@colorado.edu) are assistant professors of marketing, and Donald R. Lichtenstein (donald.lichtenstein@colorado.edu) is professor of marketing at the Leeds School of Business, University of Colorado, 419 UCB, Boulder, CO 80309. The authors acknowledge the helpful input of the editor, associate editor, and reviewers. The authors owe special thanks to Dee Warmath and also thank Andrew Long, Christina Kan, Erin Percival, Griffin Bohm, Brittany Hallett, and Bridget Leonard for research assistance, and John Lynch for comments on an earlier draft. In addition, the authors thank participants in seminars and lab groups at the University of Colorado at Boulder, Erasmus University, Ghent University, Stanford University, University of Southern California, Catholic University of Louvain, University of Michigan, and Yale University.

Vicki Morwitz served as editor, and Praveen Kopalle served as associate editor for this article.

Advance Access publication September 10, 2015

 • Vol. 42 • 2016
DOI: 10.1093/jcr/ucv047

ratings allegedly provide an almost perfect indication of product quality with little search costs (Simonson 2014, 2015; Simonson and Rosen 2014, but see Lynch 2015). As a consequence, consumers are supposedly becoming more rational decision makers, making objectively better choices, and becoming less susceptible to the influence of marketing and branding. The implications for business decision making are also profound. If these contentions are correct, businesses should be shifting resources from marketing and brand building to engineering and product development.

These conclusions rest on two key assumptions. The first assumption is that user ratings provide a good indication of product quality. The second assumption is that consumers are drawing appropriate quality inferences from user ratings. The objective of this article is to evaluate both of these assumptions. The biggest challenge in doing so is that quality is a multidimensional construct; consumers care both about objective or technical aspects of product performance (e.g., durability, reliability, safety, performance) and about more subjective aspects of the use experience (e.g., aesthetics, popularity, emotional benefits; Zeithaml 1988). Objective quality can be assessed using appropriate scientific tests conducted by experts (e.g., *Consumer Reports* scores). In contrast, subjective quality is harder to pin down because it varies across individuals and consumption contexts. For this reason, our main analyses examine the actual and perceived relationships between the average user rating and *objective* quality. We concede that consumers may consult user ratings to learn about subjective quality in addition to objective quality, and therefore the average user rating need not be a perfect indicator of objective quality to provide value to consumers. That said, we restrict our investigation to product categories that are relatively vertically differentiated (Tirole 2003), those in which alternatives can be reliably ranked according to objective standards (e.g., electronics, appliances, power tools). While products in these categories often have some subjective attributes, consumers typically care a lot about attributes that are objective (Mitra and Golder 2006; Tirunillai and Tellis 2014), and firms tout superiority on these dimensions in their advertising (Archibald, Haulman, and Moody 1983). We contend therefore that it is a meaningful and substantively important question whether the average user rating is a good indicator of objective quality and whether this squares with quality inferences that consumers draw from it.

OVERVIEW OF STUDIES AND KEY FINDINGS

This article examines empirically the actual and perceived relationships between the average user rating and objective quality. We first examine the actual relationship by analyzing a data set of 344,157 Amazon.com ratings of 1272 products in 120 product categories, which also includes quality scores from *Consumer Reports* (the most widely used indicator of objective quality in the academic literature), prices, brand image measures, and two independent sources of resale values in the used-product market. Next, we report several consumer studies designed to assess how consumers use ratings and other observable cues to form quality inferences and purchase intentions. We then compare the objective quality information that ratings actually convey to the quality inferences that consumers draw from them. This approach of comparing "ecological validity" with "cue utilization" has a long tradition in the psychology of perception, judgment, and decision making (e.g., the Lens model; Brunswik 1955; Hammond 1955).

The broad conclusion from our work is that there is a substantial disconnect between the objective quality information that user ratings actually convey and the extent to which consumers trust them as indicators of objective quality. Here is a summary of some of the key findings:

1. Average user ratings correlate poorly with *Consumer Reports* scores. Surprisingly, price is more strongly related to *Consumer Reports* scores than the average user rating. In a regression analysis with *Consumer Reports* scores as the dependent variable, the coefficient of price is almost four times that of the average user rating, and price uniquely explains 17 times as much variance in *Consumer Reports* scores as the average user rating. For two randomly chosen products, there is only a 57% chance that the product with the higher average user rating is rated higher by *Consumer Reports*. Differences in average user ratings smaller than 0.40 stars are totally unrelated to *Consumer Reports* scores such that there is only a 50% chance that the product with the higher average user rating is rated higher by *Consumer Reports*. But even when the difference is larger than one star, the item with the higher user rating is rated more favorably by *Consumer Reports* only about 65% of the time.
2. The correspondence between average user ratings and *Consumer Reports* scores depends on the number of users who have rated the product and the variability of the distribution of ratings. Averages based on small samples and distributions with high variance correspond less with *Consumer Reports* scores than averages based on large samples and distributions with low variance. However, even when sample size is high and variability low, the relationship between average user ratings and *Consumer Reports* scores is weaker than the relationship between price and *Consumer Reports* scores.
3. Average user ratings do not predict resale value in the used-product marketplace. In contrast, quality scores from *Consumer Reports* do predict resale value. We find the same results using two

independent sources of resale prices, a website that tracks prices for all products sold by third parties on the Amazon.com website and a proprietary so-called blue-book database of resale prices for digital cameras.

4. Average user ratings are influenced by price and brand image. After controlling for *Consumer Reports* scores, products have a higher user rating when they have a higher price and when they come from a brand with a premium reputation. The combined influence of these variables on the average rating is much larger than the effect of objective quality, as measured by *Consumer Reports*, explaining more than four times as much variance.
5. Consumers fail to consider these issues appropriately when forming quality inferences from user ratings and other observable cues. They place enormous weight on the average user rating as an indicator of objective quality compared to other cues. They also fail to moderate their reliance on the average user rating when sample size is insufficient. Averages based on small samples and distributions with high variance are treated the same as averages based on large samples and distributions with low variance.

THEORETICAL BACKGROUND

We are not the first to raise doubts about the value of user ratings. Several articles have voiced concerns about whether the sample of review writers is representative for the population of users. Review writers are more likely to be those that "brag" or "moan" about their product experience, resulting in a bimodal distribution of ratings for which the average does not give a good indication of the true population average (Hu, Pavlou, and Zhang 2006). There are also cross-cultural and cross-linguistic differences in the propensity to write reviews and rating extremity (De Langhe et al. 2011; Koh, Hu, and Clemons 2010). Another issue leading to nonrepresentativeness is review manipulation. Firms (or their agents) sometimes post fictitious favorable reviews for their own products and services and/or post fictitious negative reviews for the products and services of their competitors (Mayzlin, Dover, and Chevalier 2014). Moreover, many reviewers have not actually used the product (Anderson and Simester 2014), and raters that have actually used the product are influenced by previously posted ratings from other consumers and experts, creating herding effects (Jacobsen 2015; Moe and Trusov 2011; Muchnik, Aral, and Taylor 2013; Schlosser 2005). Although these findings raise general concerns about the value of user ratings, no previous research has comprehensively analyzed whether the average user rating is a good indicator of objective quality and whether the actual validity is aligned with consumer beliefs.

Convergence with *Consumer Reports* Scores

If the average user rating reflects objective quality, it should correlate positively with other measures of objective quality. We examine the extent to which average user ratings converge with *Consumer Reports* quality scores. Recognizing that even expert ratings are subject to measurement error, *Consumer Reports* scores are the most commonly used measure of objective product quality in marketing (Gerstner 1985; Hardie, Johnson, and Fader 1993; Lichtenstein and Burton 1989; Mitra and Golder 2006; Tellis and Wernerfelt 1987), as well as in psychology (Wilson and Schooler 1991) and economics (Bagwell and Riordan 1991). This is due to the impartiality and technical expertise of the organization. As noted by Tellis and Wernerfelt (1987, 244), *Consumer Reports* "is an independent body that is not allied in any way to any group of firms," and it "has a scientific approach to analyzing quality through blind laboratory studies, which in scope and consistency is unrivaled in the U.S. and in the world." This perspective is echoed by Mitra and Golder (2006, 236) who state that "several factors contribute to the objectivity of *Consumer Reports*' quality ratings including rigorous laboratory tests conducted by experts. These tests constitute one of the most elaborate quality rating systems in the world. . . . As a result, the ratings represent the most trusted objective quality information for consumers" (see also Curry and Faulds 1986; Golder, Mitra, and Moorman 2012). To our knowledge, only one article has directly examined the correspondence between user ratings and expert judgments of product quality, but this research only analyzed a single product category (Chen and Xie 2008).

One critical factor that limits the ability of the average user rating to serve as a good indicator of quality is whether it is based on a sufficient sample size. The sufficiency of the sample size depends both on the sample size itself and the variability of the distribution of ratings. Ceteris paribus, the average user rating should be more informative as sample size increases relative to variability. Unfortunately, average user ratings are often based on small samples. Moreover, variability is often high because of heterogeneity in use experience and measurement error. Users may have a fundamentally different experience or disagree in how to evaluate the experience. Alternatively, they may give a poor rating due to a bad experience with shipping, may accidentally review the wrong product, or may blame a product for a failure that is actually due to user error. Some consumers may view the purpose of product reviews differently than others. For instance, some consumers may rate purchase value (quality for the money), thereby penalizing more costly brands, whereas others may rate quality without considering price. These factors suggest that the average rating may often be based on an insufficient sample size, limiting its ability to reflect quality. We examine how convergence with

Consumer Reports scores varies as a function sample size and variability.

Ability to Predict Resale Values

High-quality products retain more of their value over time. For instance, used cars with better reliability and performance retain more of their original selling price (Ginter, Young, and Dickson 1987). Thus if average user ratings reflect objective quality, they should correlate positively with resale values. If average user ratings do not correlate with resale values, this would be evidence that they are not good measures of objective quality. We assess the ability of average user ratings to predict resale values, using the predictive ability of *Consumer Reports* as a benchmark. Because of *Consumer Reports*' technical expertise and emphasis on objective performance, we expect that *Consumer Reports* scores will have higher predictive validity for resale prices compared to average user ratings. We test this prediction via two analyses using independent data sources. We collect used prices for products in our database from an online source (camelcamelcamel.com) that reports prices for used products offered by third-party sellers on Amazon.com. We also collect blue-book prices for used products from an online data source (usedprice.com) for the largest product category in our data set (digital cameras).

The Influence of Price and Brand Image

Whereas experts like those at *Consumers Reports* have the knowledge, equipment, and time to discern objective quality through appropriate tests, consumers who post reviews and ratings typically do not. Thus it is likely that user ratings do not just reflect objective quality but also subjective quality. Extrinsic cues, such as a product's price and the reputation of the brand, are known to affect subjective evaluations of product quality (Allison and Uhl 1964; Braun 1999; Lee, Frederick, and Ariely 2006; McClure et al. 2004; Plassman et al. 2008). These variables may similarly affect average user ratings. Consumers may also engage in motivated reasoning to justify buying certain kinds of products such as those that are expensive or those made by a favored brand (Jain and Maheswaran 2000; Kunda 1990). A product may thus receive a higher rating by being more expensive or by being manufactured by a favored brand, independent of its objective quality.

These "top-down" influences on product evaluations are most pronounced when objective quality is difficult to observe (Hoch and Ha 1986). There is good reason to believe that this is often the case for vertically differentiated product categories. Product performance on important dimensions is often revealed only under exceptional circumstances. For instance, when considering a car seat, new parents would likely place a high value on crash protection, an attribute that they hope never to be in a position to evaluate. More generally, objective quality is difficult to evaluate in many categories, especially in the short time course between purchase and review posting, typically only days or weeks. In such cases, consumers are likely to draw on extrinsic cues to form their evaluations.

We examine how brand image and price setting relate to user ratings, controlling for *Consumer Reports* scores. If users are influenced by extrinsic cues when rating products, we may find a positive relationship between price and average user rating and between brand image and average user rating.

User Ratings and Consumer Quality Inferences

An obvious reason that user ratings have such a strong effect on consumer decision making and sales is via their influence on perceived quality. Given the number of potential limitations of user ratings just enumerated, the strong quality inferences that consumers presumably draw from them may not be justified. A seminal body of research on the psychology of prediction shows that people typically overweight a predictive cue when the cue is "representative" of the outcome, a phenomenon referred to by Tversky and Kahneman (1974) as the "illusion of validity." They write, "[P]eople often predict by selecting the outcome that is most representative of the input. The confidence they have in their prediction depends primarily on the degree of representativeness (that is, on the quality of the match between the selected outcome and the input) with little or no regard for the factors that limit predictive accuracy" (1126). We propose that because user ratings are highly representative of quality in the minds of consumers, they will exert a stronger effect on quality inferences than other available cues, even if those cues are actually more predictive.

The other contributor to the illusion of validity is the underweighting or complete neglect of factors that limit validity. Making a quality inference from user ratings requires intuitive statistics. Unfortunately, people are chronically poor at making statistical inferences (Kahneman and Tversky 1982). They tend to believe that the characteristics of a randomly drawn sample are very similar to the characteristics of the overall population. For instance, when judging the likelihood that one population mean is higher than another given information about sample mean, sample size, and standard deviation (SD), people are almost insensitive to sample size and SD (Obrecht, Chapman, and Gelman 2007). Findings like these suggest that consumers may jump to strong, unwarranted conclusions about quality on the basis of small sample sizes. Finally, consumers are also likely to neglect other threats to validity previously enumerated, such as the

nonrepresentativeness of the sample of review writers and the influence of price and brand image.

DO USER RATINGS REFLECT OBJECTIVE QUALITY?

Data

We visited the website of *Consumer Reports* (ConsumerReports.org) in February 2012 and extracted quality ratings for all items within all product categories where *Consumer Reports* provides these data, except for automobiles (which are not sold on Amazon.com), wine, coffee, and chocolate (which are less vertically differentiated; see pilot study later). This resulted in ratings for 3749 items across 260 product categories. To ensure that product categories were relatively homogeneous and quality ratings were comparable across items within a category, we defined product categories at the lowest level of abstraction. For example, *Consumer Reports* provides product ratings for air conditioners subcategorized by BTUs (e.g., 5000 to 6500 as opposed to 7000 to 8200). That is, brands are only rated relative to other brands in the subcategory. Thus we treated each subcategory as a separate product category. For each item for which we had a quality score from *Consumer Reports*, we searched the Amazon.com website and recorded all user ratings and the price. We were able to find selling prices and at least one Amazon.com user rating for 1651 items across 203 product categories. We further restricted the data set to products rated at least five times, and product categories with at least three products in them. The final data set consisted of 1272 products across 120 vertically differentiated product categories. See online appendix A for a list of product categories.

To verify that consumers agree that these product categories are vertically differentiated, that is, that products in these categories can be objectively ranked with respect to quality, we ran a pilot study. We paid 150 U.S. residents from Amazon Mechanical Turk $0.50 to rate 119 of the 120 categories used in our market data analysis (one category was omitted due to a programming error) in terms of whether it is possible to evaluate product quality objectively in that category. Participants read, "Some products are objectively better than others because they simply perform better. For example, a car battery that has a longer life is objectively better than one that has a shorter life. Battery life can be measured on an objective basis, that is, how long a battery lasts is not a matter of personal taste or opinion. However, for other types of products, the one that is better is a matter of individual taste. For example, one brand of potato chips is neither objectively better nor objectively worse than another brand of potato chips; it simply depends on which one the particular consumer finds more pleasurable to eat. With this difference in mind, for each of the product categories listed below, please tell us the degree to which you believe that the product category is one where one product in the category has the possibility of being objectively better than another rather than depending on the particular consumer's personal taste." For each product category, participants then responded to the following scale item: "For two different products in this product category, it is possible that one product performs better than another on objective grounds," "Strongly disagree" (1) to "Strongly agree" (5). All 119 product categories had an average rating above the scale midpoint, indicating vertical differentiation. The average rating was 3.78 of 5 ($SD = 0.17$), significantly above the scale midpoint ($t(118) = 50.30$, $p < .001$). As a reference, we also asked participants to rate 11 additional product categories (artwork, cola, jewelry boxes, wine, autobiographical books, women's perfume, chocolate cookies, men's ties, DVDs, greeting cards, and coffee) that we believed to be horizontally differentiated. The average rating for these categories was 2.53 ($SD = 0.20$), significantly below the scale midpoint ($t(10) = -7.79$, $p < .001$). None of these categories had an average rating above the scale midpoint.

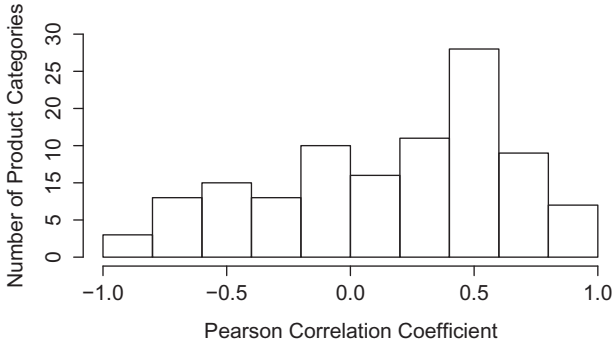
Convergence with *Consumer Reports* Scores

Simple Correlations. As a first test of the convergence between average user ratings and *Consumer Reports* scores, we computed the Pearson correlation between average user ratings and *Consumer Reports* scores for the 120 product categories in our database. These correlations are provided for each product category in online appendix A, and Figure 1 shows a histogram reflecting the distribution of these correlations. The average correlation is 0.18, and 34% of correlations are negative.

Regression Analyses. We further examined the correspondence between average user ratings and *Consumer Reports* scores for the 1272 products in our database using

FIGURE 1

DISTRIBUTION OF PEARSON CORRELATIONS BETWEEN AVERAGE USER RATINGS AND *CONSUMER REPORTS* SCORES



regression analyses. As discussed earlier, the sufficiency of the sample size should affect the ability of the average user rating to reflect quality. As a measure of the sufficiency of the sample size, we computed the standard error (SE) of the mean, or the SD divided by the square root of the sample size ($SE = SD/\sqrt{N}$). We should note that since users who rate products online are a nonprobability sample of all users of the product, we do not use the SE in any inferential manner. Rather, we use it only descriptively in that smaller SEs reflect more sufficient sample sizes. We predict an interaction between SE and average ratings, such that more sufficient sample sizes will have higher convergence with *Consumer Reports* scores. The median number of ratings for the items in our database was 50, and the average number of ratings was 271. The median SD was 1.36, and the average SD was 1.31. The median SE was 0.17, and the average SE was 0.22. Because the distribution of SEs was positively skewed, we replicated all subsequent regression analyses after log-transforming SEs. The results and conclusions remain the same.

We first regressed *Consumer Reports* scores on (1) the average user rating, (2) the SE of the average user rating, and (3) the interaction between the average user rating and the SE of the average user rating. We standardized all predictor variables by product category before analysis such that they had a mean of zero and an SD of one. Parameter estimates and confidence intervals (CIs) are shown in Table 1 (market study model A). As predicted, there was a significant interaction between the average user rating and its SE ($b = -0.06$, 95% CI, -0.12 to -0.01) such that average user ratings with higher SEs corresponded less with *Consumer Reports* scores than average user ratings with lower SEs. At the mean level of SE, *Consumer Reports* scores were significantly and positively related to average user ratings, but the effect was quite weak, consistent with the simple correlations noted earlier ($b = 0.16$, 95% CI, 0.10–0.22). Unexpectedly, the regression analysis also revealed a significant effect of SE at the mean level of average rating ($b = -0.13$, 95% CI, -0.20 to -0.07]), such that lower SEs were associated with higher *Consumer Reports* scores.

We thought this effect might be traced to the number of ratings, which has a positive effect on SE. Products with higher *Consumer Reports* scores may be more popular or be sold for a longer period of time, which would lead to a higher number of ratings. To explore this possibility, we estimated another regression model now including the number of user ratings and the SD of user ratings as predictors, in addition to the average user rating. This analysis revealed that the number of ratings was indeed positively related to *Consumer Reports* scores ($b = 0.12$, 95% CI, 0.07– 0.18) while the SD of user ratings (the other component of the SE) was not significantly related to *Consumer Reports* scores ($b = 0.06$, 95% CI, -0.01 to 0.13).

Next, we sought to benchmark the effect of average ratings on *Consumer Reports* scores to that of price. Numerous studies indicate that the correlation between price and expert ratings of objective quality is approximately between 0.20 and 0.30 (Lichtenstein and Burton 1989; Mitra and Golder 2006; Tellis and Wernerfelt 1987), and we expect to find a similar relationship strength. Including price in the model also provides a more conservative test of the hypothesis that convergence between user ratings and *Consumer Reports* scores is weak. Average user ratings may reflect purchase value to some consumers (quality – price) instead of only quality. Failing to control

TABLE 1

PARAMETER ESTIMATES (AND CONFIDENCE INTERVALS) FOR MARKET AND CONSUMER STUDIES

	Market study		Consumer studies			
	Model A	Model B	Study 2	Study 3		Study 4
Dependent variable	*Consumer Reports* quality scores	*Consumer Reports* quality scores	Perceptions of expert (*Consumer Reports*) quality scores	Perceptions of quality	Purchase likelihood	Perceptions of expert (*Consumer Reports*) quality scores
Independent variables						
Average user rating	0.16 (0.10–0.22)	0.09 (0.03–0.15)	0.34 (0.31–0.38)	0.40 (0.35–0.45)	0.35 (0.30–0.40)	0.67 (0.64–0.70)
Price		0.34 (0.28–0.39)	0.21 (0.17–0.24)	0.13 (0.08–0.17)	−0.41 (−0.46 to −0.37)	0.02 (−0.01 to 0.04)
Number of ratings			0.14 (0.10–0.18)	0.22 (0.17–0.26)	0.24 (0.19–0.28)	0.22 (0.19–0.25)
Standard error	−0.13 (−0.20 to −0.07)	−0.15 (−0.21 to −0.09)	0.00 (−0.04 to 0.04)	0.04 (−0.01 to 0.09)	0.02 (−0.03 to 0.07)	
Average user rating × standard error	−0.06 (−0.12 to −0.01)	−0.07 (−0.12 to −0.02)	0.03 (−0.01 to 0.07)	0.00 (−0.05 to 0.06)	−0.01 (−0.07 to 0.05)	
Average user rating × number of ratings						0.01 (−0.02 to 0.04)

for price may attenuate the correlation between average user ratings and quality scores from *Consumer Reports* (which measures quality, independent of price). We thus regressed *Consumer Reports* scores on (1) the average user rating, (2) the SE of the average user rating, (3) the interaction between the average user rating and the SE of the average user rating, and (4) price. Again, we standardized all predictor variables by product category before analysis, allowing us to directly compare the parameter estimates for the average user rating and price to each other. Parameter estimates and CIs are shown in Table 1 (market study model B). This analysis revealed similar results to the model without price. The interaction between average user rating and SE was again significant ($b = -0.07$, 95% CI, -0.12 to -0.02), showing that convergence between average ratings and *Consumer Reports* scores increases as SE decreases. At the mean level of SE, the average user rating was weakly but significantly related to *Consumer Reports* scores ($b = 0.09$, 95% CI, 0.03–0.15). Also the simple effect of SE at the mean level of average user rating was again significant ($b = -0.15$, 95% CI, -0.2 to -0.09). Price was not interacted with SE, so the coefficient reflects the main effect of price on *Consumer Reports* scores. This effect was significant and positive, and much stronger than the effect of average rating ($b = 0.34$, 95% CI, 0.28–0.39). The estimate for the relationship strength between price and *Consumer Reports* scores is consistent with prior estimates documented in the literature. To evaluate the relative amount of unique variance in *Consumer Reports* scores explained by price and average rating, we computed squared semipartial correlations (Cohen et al. 2003). Price uniquely explained 10.85% of the variance, 17 times more than the average user rating, which uniquely explained only 0.65%.

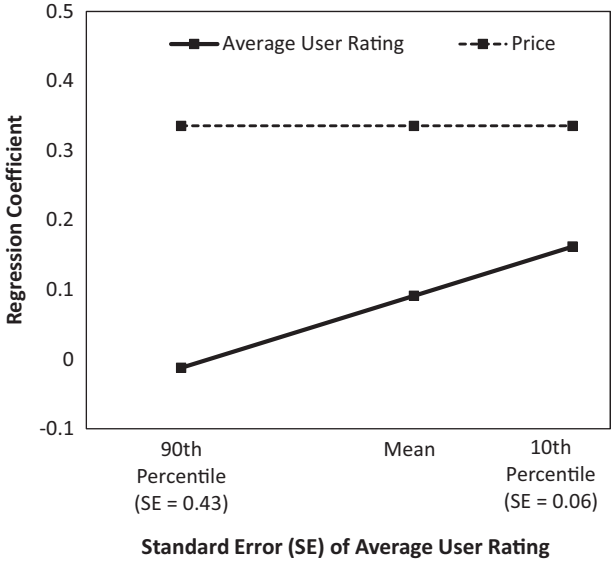
Figure 2 illustrates how the regression coefficient for the average user rating changes as a function of SE. As a reference, the chart also shows the regression coefficient for price, which is not allowed to vary as a function of SE in the regression model. At the 90th percentile of SE ($SE = 0.43$), the average user rating is unrelated to *Consumer Reports* scores. The convergence between average user ratings and *Consumer Reports* scores increases as SE decreases, but even at the 10th percentile of SE ($SE = 0.06$), the regression coefficient is still only about half that of price. In summary, price is a much better predictor of *Consumer Reports* scores than average user rating at all levels of SE.

Discussion. The regression analyses provide evidence of some degree of correspondence between average user ratings and *Consumer Reports* scores. That recognized, the correspondence is limited, in part because sample sizes are often insufficient. However, even when sample sizes are large and variability low, *Consumer Reports* scores correlate much more with price than with the average user rating. An extensive research stream has examined the correlation between price and objective quality (as measured by *Consumer Reports*). A key conclusion from this stream of research is that consumers should be cautious when inferring objective quality from price because the average price–quality correlation in the marketplace is low (typically between 0.20 and 0.30). However, consumer beliefs about the strength of the price–quality relationship tend to be inflated (Broniarczyk and Alba 1994; de Langhe et al. 2014; Gerstner 1985; Kardes et al. 2004; Lichtenstein and Burton 1989), which leads to overspending and consumer dissatisfaction (Lichtenstein, Bloch, and Black 1988; Ofir 2004). The fact that the correlation between average user ratings and *Consumer Reports* scores is so much lower suggests that an even stronger note of caution is needed when consumers infer objective quality from user ratings.

One potential objection to our conclusions is that consumers may use a different weighting scheme for valuing objective quality dimensions than *Consumer Reports*. *Consumer Reports* tests and scores products on multiple dimensions and then combines this information in some way to arrive at a composite quality score. One could argue that consumers are just as able to evaluate the quality of product dimensions as *Consumer Reports* but use a different aggregation rule, leading to a low correlation. A substantial literature in marketing (Curry and Faulds 186; Kopalle and Hoffman 1992) and in other fields such as psychology (Dawes 1979) has explored how sensitive an index derived from a weighted combination of subscores is to the weights used in the aggregation rule. The major analytical finding

FIGURE 2

CONVERGENCE BETWEEN AVERAGE USER RATINGS AND *CONSUMER REPORTS* SCORES AS A FUNCTION OF STANDARD ERROR



is that when the covariance matrix between subscores is predominantly positive, variation of weights has little effect on the composite index. The implication of this result for our research is that if product attribute covariances are predominantly positive in our product categories, we would still expect a high correlation between user ratings and *Consumer Reports* scores if consumers score product attributes similarly to *Consumer Reports* but weight them differently. Previous research in marketing has shown that covariances between product attribute quality scores are indeed predominantly positive and thus relatively insensitive to the weights assigned to dimensions when generating a composite score. Curry and Faulds (1986) found that for the vast majority of 385 product categories examined by *Test* (a German rating agency comparable to *Consumer Reports*), the covariance structure was either all positive or predominantly positive.

To evaluate whether our results are susceptible to this criticism, we supplemented our data set with attribute scores from the *Consumer Reports* website and back issues of the magazine, and ran a Monte Carlo simulation to assess how variation in the weights applied to attribute dimensions affects how correlated a judge's overall scores would be to *Consumer Reports*' overall scores. To summarize the results, similar to Curry and Faulds (1986), covariances were primarily positive (72% of covariances, averaged across categories). Consistent with this, the Monte Carlo simulation showed that variations in the weighting rule have little effect on the expected correlation. The plausible range of values for the correlation between user ratings and *Consumer Reports* scores, across categories, assuming consumers have different weights than *Consumer Reports* but score the attributes the same, is between 0.70 and 0.90. Thus attribute weighting does not explain the mismatch between user ratings and *Consumer Reports* scores. Details of the simulation and results are provided in online appendix B.

Ability to Predict Resale Values

Data. To examine whether average user ratings predict resale values, we conducted two independent analyses. First, we assessed the ability of average user ratings to predict prices in the resale marketplace for as many product categories in our database as possible. For this purpose, we augmented our database in January 2013 with used prices from the camelcamelcamel.com website that provides used prices of products sold by third parties on the Amazon.com website. The website reports the average used price over the past 50 lowest prices offered, as well as the current used price (and in the case of multiple sellers, the lowest used current price). In cases where no third-party sellers are currently selling a used version of the product, the website reports the most recent price for the used product when it was last available for sale. We conducted the analysis using the average used price over the past 50 lowest prices offered and the current used price as dependent variables. Because results are virtually identical for both dependent measures, here we only report results for the average used price. The website does not provide any information regarding the condition of the item; thus variance on this dimension is noise in the analysis. We were able to find average used prices for 1048 products across 108 product categories.

Our second analysis focuses on digital cameras, the product category in our data set with the largest number of alternatives ($N = 144$). In December 2014, we purchased a database of used prices from usedprice.com. Usedprice.com derives blue-book values from dealer surveys. The used price is calculated based on what an average store could sell the product for in 30 days or less. We were able to find used prices for 128 digital cameras in our database. Usedprice.com offers six current prices for each used camera: low and high current used retail market values, low and high current used trade-in values (mint condition), and low and high current used wholesale trade-in values (average condition). Because all six prices are highly correlated, we averaged the six values into a single used market price.

For both analyses, we assessed the ability of Amazon.com user ratings to predict used prices, using the predictive ability of *Consumer Reports* scores as a benchmark. To control for the original price of the product, we included the price of the product offered as new on Amazon.com at the time we gathered the original data set (February 2012).

Results. We standardized all variables by product category and then regressed the average used prices from camelcamelcamel.com on new prices and average user ratings. This regression revealed a significant effect of new prices ($b = 0.70$, 95% CI, 0.65–0.74), while the effect for average user ratings was just short of significance ($b = 0.04$, 95% CI, −0.003 to 0.085). New prices uniquely explained 46.8% of the variance in used price; average user ratings uniquely explained 0.2%. We then added *Consumer Reports* scores to the regression model. *Consumer Reports* scores were a highly significant predictor of used prices ($b = 0.16$, 95% CI, 0.11–0.21), uniquely explaining 2.2% of the variance. The effect of new prices remained significant ($b = 0.64$, 95% CI, 0.60–0.69), explaining 35.1% of the variance, while the effect of average user ratings was not significant ($b = 0.02$, 95% CI, −0.02 to 0.06), uniquely explaining 0.0% of the variance. We performed the same analyses for the used digital camera prices from usedprice.com.

The pattern of results was highly similar. A regression of used prices on new prices and average user ratings revealed a significant effect of new prices ($b = 0.65$, 95% CI, 0.50–0.80) but no effect for average user ratings ($b = 0.06$,

95% CI, −0.08 to 0.21). New prices uniquely explained 35.9% of the variance in used price, while average user ratings uniquely explained 0.3%. We then added *Consumer Reports* scores to the regression model. Again, *Consumer Reports* scores were a highly significant predictor of used prices ($b = 0.32$, 95% CI, 0.18– 0.47), uniquely explaining 8.4% of the variance. The effect of new prices remained significant ($b = 0.51$, 95% CI, 0.35–0.66]), explaining 18.0% of the variance, while the effect of average user ratings was not significant ($b = -0.008$; 95% CI, −0.15 to 0.13]), uniquely explaining 0.0% of the variance. Thus the totality of these results provides evidence that *Consumer Reports* scores were able to predict resale values but average user ratings were not.

DO USER RATINGS PROVIDE INFORMATION BEYOND OBJECTIVE QUALITY?

Our analyses of market data suggest that average user ratings do not converge well with *Consumer Reports* scores, even when sample sizes are large and variability is low. This could be because average user ratings are influenced by variables that influence subjective evaluations of quality, as we hypothesized in the introduction. We examine the influence of price and brand image, considered to be two of the most influential extrinsic cues for quality (Monroe and Krishnan 1985). In this analysis we regress the average user rating on these two variables while controlling for *Consumer Reports* scores. We interpret any partial effects of these variables on the average user rating as reflecting an influence of price and brand that is unrelated to objective quality.

Data

We already had selling prices in the database. In addition, we supplemented the database with brand image measures from a proprietary consumer survey conducted by a leading market research company. This survey is administered to a representative sample of U.S. consumers annually and asks multiple questions about shopping habits and attitudes toward retailers and brands across numerous product categories. We obtained data from three versions of the survey that together covered most of the product categories in our database: electronics (e.g., televisions, computers, cell phones), appliances and home improvement (e.g., blenders, refrigerators, power tools), and housewares (e.g., dishes, cookware, knives). For the brand image portion of the survey, participants were first asked to rate familiarity of all brands in the category and then were asked further questions about brand image for three brands for which their familiarity was high. All brand image questions were asked on 5 point agree/disagree Likert scales. The brand image questions differed somewhat across the three versions of the survey, so we retained data only for the 15 brand image questions that were asked in all three versions of the survey. We removed data from participants who did not complete the survey or who gave the same response to all brand image questions. We were able to realize brand image measures for 888 products representing 132 brands across 88 product categories. The data consisted of ratings from 37,953 respondents with an average of 288 sets of ratings for each brand.

For purposes of data reduction, we submitted the average value for each brand for each of the 15 questions to an unrestricted principal components analysis with a varimax rotation. This yielded three factors explaining 83% of variance in the data set. The three factors can be interpreted as brand associations related to functional benefits (seven items), emotional benefits (five items), and price (three items). While loading on separate factors, multi-item scales composed of the respective emotional and functional items were highly correlated ($r = 0.76$), leading to multicollinearity issues in subsequent regression analyses. Upon inspection of all brand image items, we found that the functional and emotional items represented what is received in the purchase (e.g., "is durable" and "is growing in popularity") while the price-related items represented sentiments related to sacrificing resources for the purchase (e.g., "is affordable"). Therefore, we repeated the principal components analysis using the a priori criterion of restricting the number of factors to two (Hair et al. 1998). The two factors accounted for 71% of variance in the data set. We interpreted the first factor to represent perceived functional and emotional benefits (12 items) and the second factor to represent perceived affordability of the brand (3 items). Because all inter-item correlations met or exceeded levels advocated in the measurement literature (see Netemeyer, Bearden, and Sharma 2003; Robinson, Shaver, and Wrightsman 1991), we averaged the respective scale items to form two brand image measures: perceived benefits ($\alpha = 0.95$) and perceived affordability ($\alpha = 0.75$). The individual scale items loading on each of the respective factors are shown in Table 2. The correlation between the two subscales was moderately negative ($r = -0.21$), suggesting that consumers see brands that provide more benefits as less affordable.

Results and Discussion

We regressed average user ratings on *Consumer Reports* scores, price, perceived brand benefits, and perceived brand affordability. We again standardized all variables by product category before analysis. The effect of selling price was significant and positive ($b = 0.10$, 95% CI, 0.03–0.17) such that more expensive products were rated more favorably. In addition, the effect of perceived brand affordability was significant and negative ($b = -0.08$, 95% CI, −0.15 to −0.01) such that products from brands that are perceived

TABLE 2

BRAND IMAGE MEASURES AND FACTOR LOADINGS

	Factor loadings	
Brand image measure	Benefits	Affordability
Has the features/benefits you want	**0.92**	−0.08
Is a brand you can trust	**0.88**	−0.25
Has high-quality products	**0.86**	−0.40
Offers real solutions for you	**0.85**	−0.03
Is easy to use	**0.82**	0.07
Has the latest trends	**0.82**	−0.05
Is durable	**0.82**	−0.34
Offers good value for the money	**0.82**	0.26
Looks good in my home	**0.80**	0.02
Offers coordinated collections of items	**0.80**	−0.07
Is growing in popularity	**0.75**	0.04
Is endorsed by celebrities	**0.32**	−0.21
Is affordable	0.00	**0.95**
Is high priced (reverse coded)	0.23	**0.83**
Has a lot of sales or special deals	−0.50	**0.80**

to be more affordable were rated less favorably. There was also a significant positive effect of perceived brand benefits ($b = 0.19$, 95% CI, 0.12–0.25) such that brands that are perceived to offer more functional and emotional benefits were rated more favorably. The total unique variance explained by these variables was 4.4%. In comparison, the unique variance explained by *Consumer Reports* scores was only 1.0% ($b = 0.11$, 95% CI, 0.04–0.18).

In sum, average user ratings are positively related to price, both at the product level (i.e., the effect of selling price) and at the brand level (i.e., the effect of a brand's perceived affordability). Surprisingly, consumers do not penalize higher priced items in their ratings. On the contrary, holding *Consumer Reports* scores constant, consumers rate products with higher prices more favorably. Brands that have a better reputation for offering benefits also obtain higher ratings. The combined effects of price and brand image are much larger than the effect of *Consumer Reports* scores.

We believe the most likely interpretation of these results is that brand image and price influence ratings. However the data are correlational, and other interpretations are possible. For instance, one alternative interpretation for the positive effect of price is that Amazon.com raises/lowers their prices in response to user ratings. While we are aware that Amazon.com sets prices based on individual level data that relates to the consumer's price sensitivity (e.g., the consumer's previous purchase history or the browser the consumer is using; see "Personalising Online Prices," 2012), we are unaware of any source that has alleged that Amazon.com adapts prices based on user ratings. Nevertheless, in order to gain some insight into this issue we collected Amazon.com prices for the brands in our data set at three additional points in time (September 22, 2012, November 22, 2012, and January 22, 2013; the main data set was collected on February 14, 2012). If user ratings influence prices, we would expect to find a positive correlation between these ratings and subsequent price changes. That is, higher ratings at time 1 (i.e., February 14, 2012) should be positively related to price changes from time 1 to time 2 (i.e., the difference in price between any of these three additional times and the price on February 14, 2012). Thus we calculated three price changes and found they were not significantly related to average user ratings on February 14, 2012 ($r_{sep} = .01$, $p > .87$; $r_{nov} = .04$, $p > .35$; $r_{jan} = -.01$, $p > .74$), which is inconsistent with the reverse causality argument.

Another potential explanation for the results is that there could be unobserved variation in objective quality that is not captured by *Consumer Reports* but is captured by price and brand image. It is commonly assumed that this is not the case, for instance in the literature on price–quality relationships and consumer learning about quality more generally (Bagwell and Riordan 1991; Curry and Faulds 1986; Erdem et al. 2008; Gerstner 1985; Hardie et al.1993; Lichtenstein and Burton 1989; Mitra and Golder 2006; Tellis and Wernerfelt 1987; Wilson and Schooler 1991). Moreover, the causal interpretation is parsimonious and consistent with a great deal of previous research showing that price and brand are powerful extrinsic cues for quality (Monroe and Krishnan 1985; Rao and Monroe 1989). From this perspective, our findings should not be surprising.

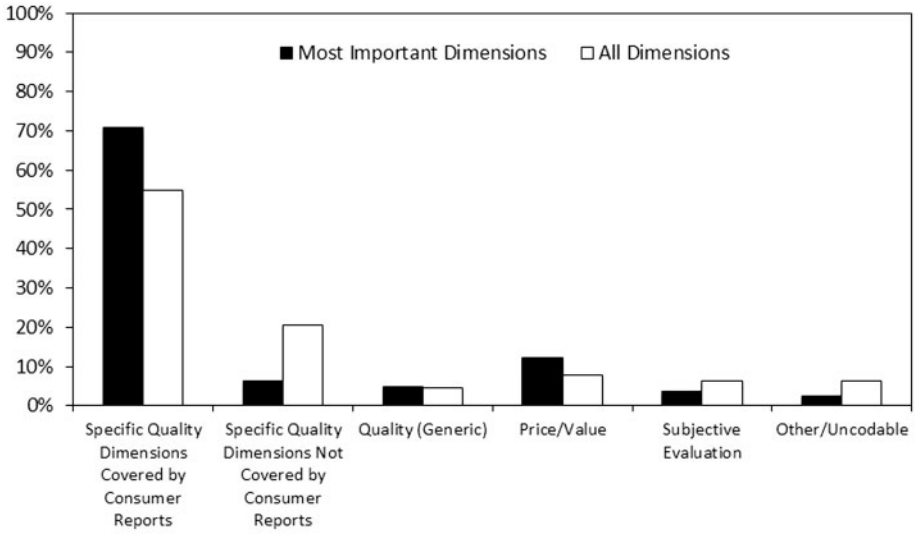
In addition to price and brand image, it is possible that user ratings also reflect other information that is not revealed through objective testing. For example, consumers may rate aesthetic aspects of a product, something that is not considered as a dimension of objective quality, but a dimension of quality that consumers value nonetheless. Also, user evaluations are typically posted shortly after purchase. These initial impressions may be based on variables that are unrelated to objective quality. This is a promising avenue for future research.

CONSUMER STUDIES

Our analyses of various secondary data sources indicate that average user ratings from Amazon.com do not converge well with *Consumer Reports* scores, are often based on insufficient sample sizes, fail to predict prices in the used-product marketplace, and are influenced by price and brand image. These analyses suggest that the average user rating lacks validity as a measure of objective quality. One potential objection to our analyses of market data is that consumers may not view user ratings as purporting to capture objective quality. Consumers might not care whether user ratings converge with *Consumer Reports* scores or whether they predict resale values. Instead consumers may

FIGURE 3

CONSUMER STUDY 1: WHY DO CONSUMERS CONSULT USER RATINGS AND REVIEWS?



believe that user ratings are meant to capture other kinds of information, like subjective aspects of the use experience, product aesthetics, or other dimensions that are not amenable to objective tests. We undertook a series of controlled studies to examine the extent to which consumers rely on the average user rating as a cue for objective quality. We summarize the main findings of these studies here and provide the methodological details and results in online appendix C.

In study 1, we asked consumers to list reasons why they consulted online ratings and reviews for a subset of product categories in our database. We also asked them to indicate the reason that was most important to them. Objective dimensions covered by *Consumer Reports* were by far the most common and most important reason (see Figure 3). Another common reason was to learn about the price or value of a product. Some consumers reported consulting user ratings and reviews to learn about more subjective evaluations but much less frequently. These results suggest that consumers consult user ratings for vertically differentiated product categories primarily to learn about technical dimensions of quality that are amenable to objective tests and are covered by *Consumer Reports*.

The goal of study 2 was to quantify consumers' reliance on the average user rating as a cue for quality and compare it to reliance on other cues for quality. We asked consumers to search for pairs of products on Amazon.com, inspect the product web pages, and then to judge which product they thought *Consumer Reports* would rate higher on a scale from 1 (product A would be rated as higher quality) to 10 (product B would be rated as higher quality). To avoid any demand effects, we designed the search and rating task to be as realistic as possible, and we gave participants no training and minimal instructions. Because the products vary naturally in terms of average user ratings and prices, we were able to test the relative influence of differences in average user ratings and differences in prices on quality judgments. We also examined the extent to which consumers used the number of user ratings as a direct cue for quality. The number of user ratings is significantly related to *Consumer Reports* scores (see earlier). Moreover, retailers frequently use promotional phrases such as "Over 10,000 Sold" because consumers may gain confidence about the quality of a product simply by knowing that many other consumers have purchased the product ("social proof"; Cialdini 2001). A large number of ratings may also indicate that the product has staying power in the market, another indication of quality. Thus it is plausible that consumers believe that products with more ratings have higher quality than products with fewer ratings.

For each product pair, we computed the difference between product A and product B in average user rating, number of user ratings, and price. We collected this data from the Amazon.com website right before launching the study. It is important to note that while we collected these three variables from the respective product web pages prior to the study, participants were exposed to the full array of information on the product web pages, thereby enhancing external validity. To measure the extent to which the sample sizes for two products in a pair were sufficiently large for the difference in average user ratings to be informative, we computed the Satterthwaite approximation for the pooled SE (hereafter referred to as "pooled SE"), which is a function of the sample sizes and the variation in user

ratings of products A and B ($SE_{Pooled} = \sqrt{[(VAR_A/N_A) + (VAR_B/N_B)]}$). A higher pooled SE indicates that sample sizes are less sufficient.

We regressed consumers' judgments of quality on (1) the difference in average user ratings, (2) the pooled SE of the difference in average user ratings, (3) the interaction between the difference in average user ratings and the pooled SE of the difference in average user ratings, (4) the difference in the number of user ratings, and (5) the difference in prices. Quality judgments were more strongly related to differences in average user ratings than to differences in prices and differences in the number of ratings. Average user ratings uniquely explained 10.98% of variance in quality judgments, more than two times as much a price that uniquely explained 4.46%, and more than five times as much as the number of ratings that uniquely explained 2.05%. Moreover, reliance on the difference in average user ratings was not moderated by the SE of the difference in average user ratings. Participants did not weigh differences in average user ratings based on sufficient sample sizes more than average user ratings based on insufficient sample sizes when judging quality. Regression results for this study, as well as consumer studies 3 and 4 (described later), are provided in Table 1.

To test the robustness of these results, we ran two additional studies similar to study 2. Study 3 used a correlational design, as in study 2, but we used a generic quality measure (rather than specifying *Consumer Reports* quality). We asked respondents to copy the values of the relevant cues to a table before judging quality, and we added a purchase intention question. The fourth study was similar, but we used a true experimental design, where we orthogonally manipulated the average rating, the price, and the number of ratings. Results were very consistent across studies 2, 3, and 4. First, consumers relied most heavily on average user ratings, which was true regardless of whether quality was defined as *Consumer Reports* quality or generically. Second, consumers did not moderate their reliance on average user ratings depending on whether sample size was sufficient or not. Third, in two of the three studies, consumers also relied on price but much less so than on average user ratings. Finally, consumers did use the number of ratings as a direct indicator of quality.

GENERAL DISCUSSION

Our analyses of market data together with the consumer studies suggests a substantial mismatch between the objective quality information that user ratings actually convey and the quality inferences that consumers draw. In the marketplace, price is the best predictor of objective quality, explaining 17 times as much variance in *Consumer Reports* scores. In contrast, the average user rating is weighted most heavily by consumers, explaining more than two times as much variance in quality judgments as price. Price has been identified in the consumer research literature as one of the most commonly used cues for quality (Rao and Monroe 1989). Consumer advocates frequently warn consumers not to assume that "they will get what they pay for," yet we are unaware of similar advice with regard to user ratings. Moreover, although average user ratings correspond less with actual *Consumer Reports* scores when sample sizes are insufficient, consumers do not take this into account when making quality inferences.

Recommendations for Consumers

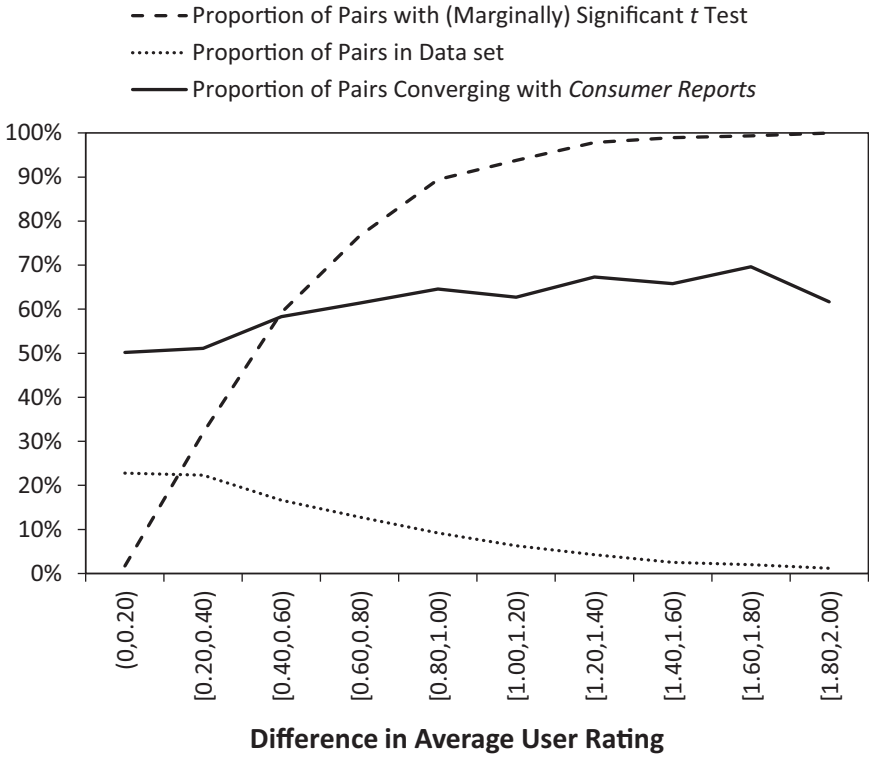
Our findings suggest that the objective quality information available in average user ratings is much weaker than what consumers believe. This evidence comes from interpretation of regression coefficients, which may not provide a good intuition for the effect sizes at issue. In this section we attempt to provide more intuitive benchmarks by presenting an analysis based on pairwise comparisons of products in our database. Consider a consumer who is trying to choose between two products in a category and observes the distribution of user ratings for each product. We address two questions: First, upon observing that one product has a higher average rating than the other, how confident can the consumer be that it also has a higher *Consumer Reports* score? Second, independent of *Consumer Reports* scores, how often are sample sizes sufficient to discriminate between the averages of the two distributions?

To address these two questions we determined all pairwise comparisons of products for each product category in our data set. This resulted in 15,552 pairs of products (after excluding pairs for which items have identical quality scores and/or identical average user ratings). We binned pairs according to the absolute magnitude of the difference (using a bin width of 0.20 stars) and, for each of the bins, calculated the proportion of times the item with the higher average user rating received a higher quality score from *Consumer Reports*. These proportions are indicated by the solid line in Figure 4. Very few comparisons had rating differences larger than two stars, so the data are only shown for differences between zero and two stars, which accounts for approximately 95% of the database. Averaging across all comparisons, the correspondence between the average user rating and *Consumer Reports* scores is only 57%. When the difference in user ratings is smaller than 0.40 stars, correspondence is at chance (50%). This percentage increases as the difference in user rating grows larger, but the increase is modest and correspondence never exceeds 70%.

A key result from the consumer studies is that consumers do not moderate their quality inferences as a function of sample size and variability of ratings. This is a problem because average user ratings based on insufficient sample sizes have no correspondence with *Consumer*

FIGURE 4

CONVERGENCE BETWEEN AVERAGE USER RATINGS AND *CONSUMER REPORTS* SCORES (PAIRWISE ANALYSIS)



Reports scores. An important rule for any consumer evaluating products based on the average user rating is not to jump to a conclusion about relative quality if the difference in averages could easily be due to chance and not due to a true difference in the average user experience. To evaluate how often sample sizes are sufficient to discriminate two average user ratings, we conducted independent samples *t* tests (assuming unequal variances) for each of the 15,552 product pairs. The *t* test evaluates the probability of obtaining a difference in average ratings this large, if in fact the two sets of ratings were sampled from a parent distribution with the same average. Prior to reporting results of this analysis, an important caveat is in order regarding our use of *t*-test analyses for addressing this issue. We noted in the introduction that the people who rate products are a nonrepresentative sample of the population of users. These *t* tests reflect what a consumer can infer about this population of review writers, not the overall population of users. The statistics based on the *t*-test analysis may not perfectly reflect whether a difference is likely to exist in the overall population of users. However, given that these biased samples are all the consumer has to work with, the *t* test provides a reasonable evaluation of whether a difference in average ratings is likely to reflect a meaningful difference in use experience.

With this caveat noted, the difference between average user ratings was at least marginally significant ($p < .10$) for 52% of pairs but nonsignificant ($p > .10$) for 48% of pairs. Thus even using a liberal criterion of $p < .10$ for assessing significance, approximately half the time a comparison between two average ratings does not clearly indicate a true difference in the average use experience. Statistical significance depends on the magnitude of the difference in average user ratings. As the difference grows larger, a smaller sample size will suffice. Thus larger star differences should be more likely to result in significant *t* tests. This is indeed what we observe, as indicated by the dashed line in Figure 4. When the difference in average user ratings is smaller than 0.20 stars, there is only 2% chance that it is statistically significant. As the difference in average user ratings grows larger to 0.40 stars, this likelihood increases to 32%. Although differences larger than one star are almost always statistically significant (97%), differences of this magnitude are relatively rare (16% of comparisons). This can be seen from the dotted line in Figure 4 that shows the proportion of product pairs in each bin.

In light of these results, how should consumers change their behavior? User ratings may have value in two ways. First, they do correspond with objective quality scores somewhat. Although the relationship is weak, it is

substantially stronger when sample sizes are sufficient. Consumers should avoid jumping to quality judgments based on insufficient sample sizes. When sample sizes are sufficient, consumers can learn something about objective quality, but they should realize the information is far from perfect and base their quality judgment on additional sources of evidence.

Second, our findings showed that ratings correlate positively with price and brand image, controlling for *Consumer Reports* scores, and we know these variables can positively influence the consumption experience (Plassman et al. 2008). In light of this, when the average rating is based on a sufficiently large sample size, but contradicts the evaluations of expert testers like *Consumer Reports*, a consumer needs to ask what she wants to optimize. If she wants to optimize performance on technical dimensions and resale value, she should follow the experts. If she wants to optimize short-term consumption utility, she may be better off following the average user rating, although we offer this possibility very tentatively. More research is needed before reaching this conclusion.

Limitations and Future Research

One limitation of our analyses is that we only analyzed quantitative star ratings while not considering narrative reviews. There is recent evidence that narrative reviews do contain useful information about product quality (Tirunillai and Tellis 2014) and that consumers consult them (Chevalier and Mayzlin 2006). Using textual analysis, Tirunillai and Tellis (2014) found that narrative reviews cover many of the same dimensions as *Consumer Reports* and that the valence of the vocabulary used to describe performance in narrative reviews correlates with *Consumer Reports* scores. However, Tirunillai and Tellis (2014) rely on an advanced statistical approach to analyze all reviews in an unbiased way. There is reason to doubt that consumers can extract this quality information from the narrative reviews. Instead of processing all reviews in a balanced way, consumers most likely rely on a limited subset of reviews, those that are most recent, vivid, extreme, emotional, and concrete (Reyes, Thompson, and Bower 1980). These reviews are not necessarily most diagnostic of product quality. To give just one example, the review ranked as most helpful at Amazon.com for the Britax Frontier Booster Car Seat is titled "Saved both of my girls' lives." It was written by "luckymom" who recently experienced a horrible accident in which both of her daughters walked away with minor cuts and bruises. The mother completely attributes the well-being of her children to the quality of the Britax car seat. Although prospective car seat buyers perceive this review to be highly informative, from a scientific point of view it should in fact be discounted because the data point was obtained in an "experiment" without a control group. Anecdotally, we have been told by several consumers that they read only the most negative reviews prior to making a purchase decision in order to gauge potential downsides of purchasing. Future research might look at how often consumers read narratives, how they integrate the narrative information with the quantitative ratings, how they choose which narratives to read, and whether the narratives help or hinder consumers' quality inferences. Our results show that whatever objective quality information is contained in the narrative reviews is not reflected very well in the average user rating. This squares with research by Tirunillai and Tellis (2012) showing that text mining of narrative reviews can be used to predict stock market performance in some cases, but the average user ratings are not predictive.

A second limitation of our data is that it does not cover the full range of products and services for which people consult online user ratings. We restricted our analyses to vertically differentiated product categories because it is well accepted that quality can be defined objectively in these categories and measured by experts. But online ratings are also pervasive in the evaluation of more experiential products like alcoholic beverages (e.g., winespectator.com) and services like restaurants (e.g., Yelp.com), hotels (e.g., tripadvisor.com), and contractors (e.g., angieslist.com), and recent research shows that consumers do indeed rely on user ratings for experiential purchases, although less so than for material purchases (Dai, Chan, and Mogilner 2014). A general concern with ratings for taste-based or horizontally differentiated goods is that learning about the average taste may not be very useful because taste is heterogeneous. One way to get around this issue, which some websites are doing (e.g., Netflix.com), is to provide a tailored average rating, which weighs certain ratings more than others (e.g., those by users deemed similar to the consumer based on transaction history).

The Role of Marketing in the New Information Environment

We began the article by describing an emerging debate in the consumer behavior literature pertaining to the large-scale implications of changes in the information environment for consumer and business decision making. Simonson and Rosen (2014) argue that we are entering an age of almost perfect information, allowing consumers to make more informed choices and be influenced less by marketers. Although we have reached a starkly different conclusion with respect to the validity of user ratings and the appropriateness of consumers' quality inferences based on these ratings, we would also like to highlight an area of agreement. We agree that the consumer information environment has changed dramatically and that these changes are having pervasive effects on consumer behavior. We are also sympathetic to the possibility that the *direct* influence of marketing may be waning. For instance, the price–quality

heuristic is one of the most studied phenomena in consumer behavior and yet, price is overshadowed as a cue to quality when user ratings are also available (see consumer studies 2, 3 and 4). This suggests that the findings from this literature need to be revisited given the rise of online shopping. More generally, many traditional consumer research topics need to be updated. Thus we strongly support the call by Simonson (2015) and others that consumer researchers start tackling issues pertaining to how consumer behavior is changing in the new information environment.

Although we agree in broad terms about these effects, we disagree on the specific claims. For the vertically differentiated product categories we have studied, user ratings are far from conveying nearly perfect information about objective quality. Consumers do not make appropriate quality inferences from ratings, instead jumping to strong, unjustifiable conclusions about quality while underutilizing other cues like price. Moreover, user ratings seem to be colored by brand image, suggesting that a new, *indirect* route of marketing influence is emerging; brand image influences consumers through their effect on user ratings. Thus while the direct influence of marketing may be waning due to the proliferation of new sources of information, this does not protect consumers from marketing influence. In fact, this indirect route might be more insidious in the sense that traditional marketing appeals trigger persuasion knowledge (Friestad and Wright 1994) while user ratings do not.

We conclude that although the information environment is changing, the psychological processes that lead consumers to give higher evaluations to premium brands, engage in motivated reasoning when reviewing a product, ignore sample size when making inferences or fall victim to illusions of validity, remain the same. In other words, imperfect people stand in the way of the age of perfect information.

DATA COLLECTION INFORMATION

The market data were collected according to procedures described in the article. The data for the pilot study used to provide evidence that the product categories are perceived as relatively vertically differentiated were collected by a research assistant under supervision of the authors. The camelcamelcamel.com data set was scraped by a third-party contractor according to specifications of the authors. The usedprice.com data set was collected by a research assistant under supervision of the authors. Brand perception measures were provided to the authors by a major marketing research firm. Data for consumer studies 1 to 4 (reported in detail in online appendix C) were collected by a research assistant under supervision of the authors. All data were analyzed by all authors.

REFERENCES

Aaker, David A. and Robert Jacobson (1994), "The Financial Information Content of Perceived Quality," *Journal of Marketing Research*, 31 (May), 191–201.

Allison, Ralph I. and Kenneth P. Uhl (1964), "Influence of Beer Brand Identification on Taste Perception," *Journal of Marketing Research*, 1 (August), 36–39.

Anderson, Eric and Duncan Simester (2014), "Reviews without a Purchase: Low Ratings, Loyal Customers and Deception," *Journal of Marketing Research*, 51 (3), 249–69.

Archibald, Robert B., Clyde A. Haulman, and Carlisle E. Moody Jr. (1983), "Quality, Price, and Published Quality Ratings," *Journal of Consumer Research*, 9 (March), 347–55.

Bagwell, Kyle and Michael H. Riordan (1991), "High and Declining Prices Signal Product Quality," *American Economic Review*, 81 (March), 224–39.

Bolton, Ruth N. and James H. Drew (1991), "A Multistage Model of Customers' Assessments of Service Quality and Value," *Journal of Consumer Research*, 17 (March), 375–84.

Braun, Kathryn A. (1999), "Postexperience Advertising Effects on Consumer Memory," *Journal of Consumer Research*, 25 (March), 319–34.

Broniarczyk, Susan M. and Joseph W. Alba (1994), "Theory versus Data in Prediction and Correlation Tasks," *Organizational Behavior and Human Decision Processes*, 57, 117–39.

Brunswik, Egon (1955), "Representative Design and Probabilistic Theory in a Functional Psychology," *Psychological Review*, 62 (3), 193–217.

Chen, Yubo and Jinhong Xie (2008), "Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix," *Marketing Science*, 54 (March), 477–91.

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

Chintagunta, Pradeep K., Shyam Gopinath, and Sriram Venkataraman (2010), "The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets, *Marketing Science*, 29 (September-October), 944–57.

Cialdini, Robert B. (2001), *Influence: Science and Practice*, Needham Heights, MA: Allyn & Bacon.

Cohen, Jacob, Patricia Cohen, Stephen G. West, and Leona S. Aiken (2003), *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, New York: Routledge.

Curry, David J. and David J. Faulds (1986), "Indexing Product Quality: Issues, Theory, and Results," *Journal of Consumer Research*, 43 (June), 134–45.

Dai, Hengchen, Cindy Chan, and Cassie Mogilner (2014), "Don't Tell Me What to Do! People Rely Less on Consumer Reviews for Experiential than Material Purchases," working paper, The Wharton School.

Dawes, Robyn M. (1979), "The Robust Beauty of Improper Linear Models in Decision Making," *American Psychologist*, 34 (7), 571–82.

De Langhe, Bart, Stijn M. J. van Osselaer, Stefano Puntoni, and Ann L. McGill (2014), "Fooled by Heteroscedastic Randomness: Local Consistency Breeds Extremity in Price-Based Quality Inferences," *Journal of Consumer Research*, 41 (December), 978–94.

De Langhe, Bart, Stefano Puntoni, Stijn M. J. van Osselaer, and Daniel Fernandes (2011), "The Anchor Contraction Effect in International Marketing Research," *Journal of Marketing Research*, 48 (April), 366–80.

Erdem, Tulin, Michael P. Keane, and Baohong Sun (2008), "A Dynamic Model of Brand Choice When Price and Advertising Signal Product Quality," *Marketing Science*, 27, 1111–25.

Floyd, Kristopher, Ryan Freling, Saad Alhoqail, Hyun Young Cho, and Traci Freling (2014), "How Online Product Reviews Affect Retail Sales: A Meta-analysis," *Journal of Retailing*, 90 (June), 217–32.

Friestad, Marian and Peter Wright (1994), "The Persuasion Knowledge Model: How People Cope with Persuasion Attempts," *Journal of Consumer Research*, 21 (1), 1–31.

Gerstner, Eitan (1985), "Do Higher Prices Signal Higher Quality," *Journal of Marketing Research*, 22 (May), 209–15.

Ginter, James L., Murray A. Young, and Peter R. Dickson (1987), "A Market Efficiency Study of Used Car Reliability and Prices," *Journal of Consumer Affairs*, 21 (Winter), 258–76.

Golder, Peter N., Debanjan Mitra, and Christine Moorman (2012), "What Is Quality? An Integrative Framework of Processes and States," *Journal of Marketing*, 76 (July), 1–23.

Hammond, Kenneth R. (1955), "Probabilistic Functioning and the Clinical Method," *Psychological Review*, 62 (4), 255–62.

Hair, Joseph F. Jr., Rolph E. Anderson, Ronald L. Tatham, and William C. Black (1998), *Multivariate Data Analysis*, Upper Saddle River, NJ: Prentice Hall.

Hardie, Bruce G. S., Eric J. Johnson, and Peter S. Fader (1993), "Modeling Loss Aversion and Reference Dependence Effects on Brand Choice," *Marketing Science*, 12 (4), 378–94.

Hoch, Stephen J. and Young-Won Ha (1986), "Consumer Learning: Advertising and the Ambiguity of Product Experience," *Journal of Consumer Research*, 13 (September), 221–33.

Hu, Nan, Paul A. Pavlou, and Jennifer Zhang (2006), "Can Online Reviews Reveal a Product's True Quality? Empirical Findings and Analytical Modeling of Online Word-of-Mouth Communication." *Proceeding of the 7th ACM Conference on Electronic Commerce*, (EC'06, June 11-15), 324–30.

Jacobsen, Grant D. (2015), "Consumers, Experts, and Online Product Evaluations: Evidence from the Brewing Industry," *Journal of Public Economics*, 126, 114–23.

Jain, Shailendra Pratap and Durairaj Maheswaran (2000), "Motivated Reasoning: A Depth-Of-Processing Perspective," *Journal of Consumer Research*, 26 (March), 358–71.

Kahneman, Daniel and Amos Tversky (1982), "On the Study of Statistical Intuitions," *Cognition*, 11, 123–41.

Kardes, Frank R., Maria L. Cronley, James J. Kellaris, and Steven S. Posavac (2004), "The Role of Selective Information Processing in Price-Quality Inference," *Journal of Consumer Research*, 31 (September), 368–74.

Koh, Noi Sian, Nan Hu, and Eric K. Clemons (2010), "Do Online Reviews Reflect a Product's True Perceived Quality? An Investigation of Online Movie Reviews Across Cultures," *Electronic Commerce Research and Applications*, 9, 374–85.

Kopalle, Praveen K. and Donna L. Hoffman (1992), "Generalizing the Sensitivity Conditions in an Overall Index of Product Quality," *Journal of Consumer Research*, 18 (4), 530–35.

Kunda, Ziva (1990), "The Case for Motivated Reasoning," *Psychological Bulletin*, 108, 480–98.

Lee, Leonard, Shane Frederick, and Dan Ariely (2006), "Try It, You'll Like It: The Influence of Expectation, Consumption, and Revelation on Preferences for Beer," *Psychological Science*, 17, 1054–58.

Lichtenstein, Donald R., Peter H. Bloch, and William C. Black (1988), "Correlates of Price Acceptability," *Journal of Consumer Research*, 15 (September), 243–52.

Lichtenstein, Donald R. and Scot Burton (1989), "The Relationship Between Perceived and Objective Price-Quality," *Journal of Marketing Research*, 26 (November), 429–43.

Loechner, Jack (2013), "Consumer Review Said to Be THE Most Powerful Purchase Influence," *Research Brief from the Center for Media Research*, http://www.mediapost.com/publications/article/190935/consumer-review-said-to-be-the-most-powerful-purch.html#axzz2Mgmt90tc.

Luca, Michael (2011), "Reviews, Reputation, and Revenue: The Case of Yelp.com," Working Paper 12-016, Harvard Business School.

Lynch, John G. Jr. (2015), "Mission Creep, Mission Impossible, or Mission of Honor? Consumer Behavior BDT Research in an Internet Age," *Journal of Marketing Behavior*, 1, 37–52.

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier (2014), "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, 104 (8), 2421–55.

McClure, Samuel M., Jian Li, Damon Tomlin, Kim S. Cypert, Latane M. Montague, and P. Read Montague (2004), "Neural Correlates of Behavioral Preference for Culturally Familiar Drinks," *Neuron*, 44 (2), 379–87.

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.

Moe, Wendy W. and Michael Trusov (2011), "The Value of Social Dynamics in Online Product Ratings Forums," *Journal of Marketing Research*, 49 (June), 444–56.

Monroe, Kent B. and R. Krishnan (1985), "The Effect of Price on Subjective Product Evaluations," in *Perceived Quality*, ed. Jack Jacoby and Jerry Olson, Lexington, MA: Lexington Books, 209–32.

Muchnik, Lev, Sinan Aral, and Sean J. Taylor (2013), *"Social Influence Bias: A Randomized Experiment," Science*, 341 (August 9), 647–51.

Netemeyer, Richard G., William O. Bearden, and Subhash Sharma (2003), *Scale Development in the Social Sciences: Issues and Applications*, Palo Alto, CA: Sage.

Obrecht, Natalie, Gretchen B. Chapman, and Rochel Gelman (2007), "Intuitive t-tests: Lay Use of Statistical Information," *Psychological Bulletin & Review*, 14 (6), 1147–52.

Ofir, Chezy (2004), "Reexamining Latitude of Price Acceptability and Price Thresholds: Predicting Basic Consumer Reaction to Price," *Journal of Consumer Research*, 30 (March), 612–21.

"Personalising Online Prices, How Deep Are Your Pockets? Businesses Are Offered Software That Spots Which Consumers Will Pay More" (2012), *The Economist,* http://www.economist.com/node/21557798.

Plassman, Hilke, John O'Doherty, Baba Shiv, and Antonio Rangel (2008), "Marketing Actions Can Modulate Neural Representations of Experienced Pleasantness," *National Academy of Sciences of the USA*, 105 (January 22), 1050–54.

Rao, Akshay R. and Kent B. Monroe (1989), "The Effect of Price, Brand Name, and Store Name on Buyers' Perceptions of

Product Quality: An Integrative Review," *Journal of Marketing Research*, August (26), 351-57.

Reyes, Robert M., William C. Thompson, and Gordon Bower (1980), "Judgmental Biases Resulting from Differing Availabilities of Arguments," *Journal of Personality and Social Psychology*, 39, 2–11.

Robinson, John P., Phillip R. Shaver, and Lawrence S. Wrightsman (1991), "Criteria for Scale Selection and Evaluation," in *Measures of Personality and Social Psychological Attitudes*, ed. J. P. Robinson, P. R. Shaver, and L. S. Wrightsman, San Diego, CA: Academic Press, 1–15.

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.

Schlosser, Ann (2005), "Posting Versus Lurking: Communicating in a Multiple Audience Context," *Journal of Consumer Research*, 32 (September), 260–65.

Simonson, Itamar (2014), "What Really Influences Customers in the Age of Nearly Perfect Information?" Marketing Science Institute Webinar, August 14, https://www.msi.org/conferences/what-really-influences-customers-in-the-age-of-nearly-perfect-information/#/speakers.

——. (2015), "Mission (Largely) Accomplished: What's Next for Consumer BDT-JDM Researchers," *Journal of Marketing Behavior*, 1, 9–35.

Simonson, Itamar and Emanuel Rosen (2014), *Absolute Value*, New York: HarperCollins.

Tellis, Gerard J. and Birger Wernerfelt (1987), "Competitive Price and Quality Under Asymmetric Information," *Marketing Science*, 6 (Summer), 240–53.

Tirole, Jean (2003), *The Theory of Industrial Organization*, Cambridge, MA: MIT Press.

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

——. (2014), "Mining Meaning from Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation," *Journal of Marketing Research*, 51 (4), 463–79.

Tversky, Amos, and Daniel Kahneman (1974), "Judgment Under Uncertainty: Heuristics and Biases," *Science* 185, 1124–31.

Wilson, Timothy D. and Jonathan W. Schooler (1991), "Thinking Too Much: Introspection Can Reduce the Quality of Preferences and Decisions," *Journal of Personality and Social Psychology*, 60 (February), 181–92.

Zeithaml, Valarie A. (1988), "Consumer Perceptions of Price, Quality, and Value: A Means-End Model and Synthesis of Evidence," *Journal of Marketing*, 52 (July), 2–22.