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 that 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

© The Author 2015. Published by Oxford University Press on behalf of Journal of Consumer Research, Inc.
All rights reserved. For permissions, please e-mail: journals.permissions@oup.com • 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.

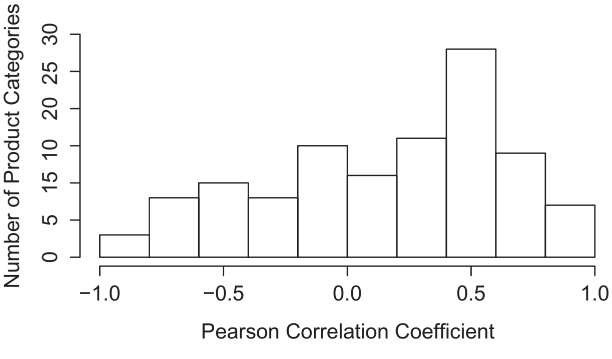
Convergence with *Consumer Reports* Scores

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%.

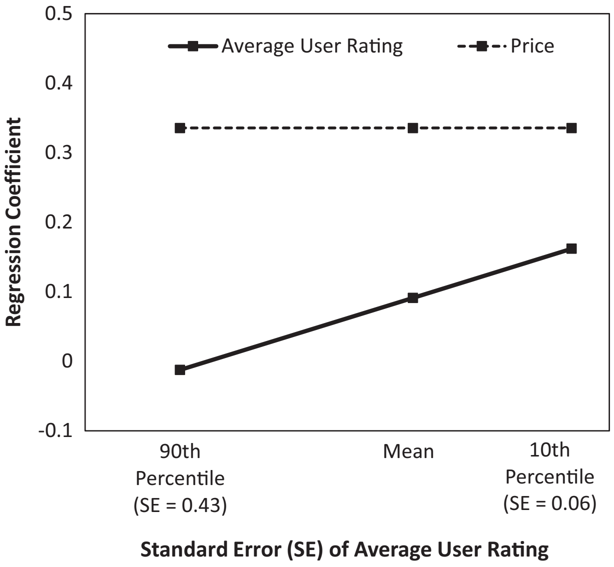
Figure 2 illustrates how the regression coefficient for the average user rating changes as a function of SE. As a reference, the chart also shows the regression coefficient for price, which is not allowed to vary as a function of SE in the regression model. At the 90th percentile of SE (SE = 0.43), the 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.

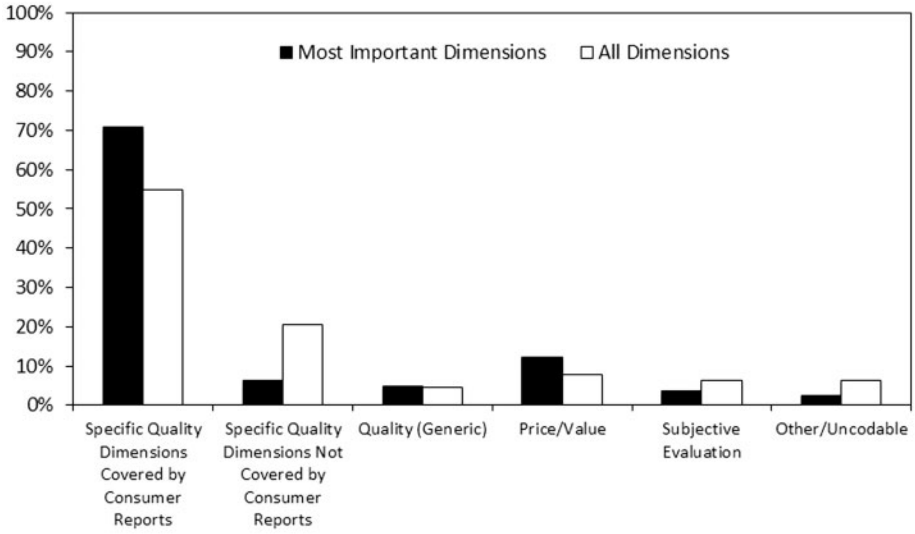
In addition to price and brand image, it is possible that user ratings also reflect other information that is not revealed through objective testing. For example, consumers may rate aesthetic aspects of a product, something that is not considered as a dimension of objective quality, but a dimension of 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 ($\mathrm{SE}_{\mathrm{Pooled}} = \sqrt{[(VAR_{\mathrm{A}}/N_{\mathrm{A}}) + (VAR_{\mathrm{B}}/N_{\mathrm{B}})]}$). 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

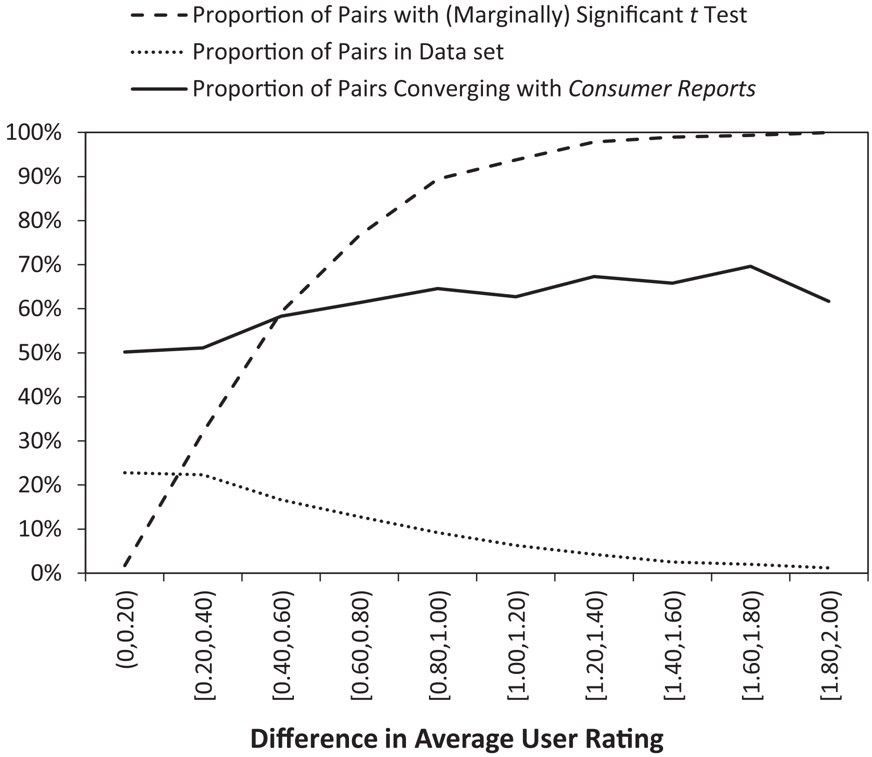
Our findings suggest that the objective quality information available in average user ratings is much weaker than what consumers believe. This evidence comes from interpretation of regression coefficients, which may not provide a good intuition for the effect sizes at issue. In this section we attempt 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.

Commentary on de Langhe, Fernbach, and Lichtenstein
Amazonian Forests and Trees: Multiplicity and Objectivity in Studies of Online Consumer-Generated Ratings and Reviews

ROBERT V. KOZINETS

Consumer-generated ratings and reviews play an important role in people's experiences of online search and shopping. This article applauds and extends the thought-provoking response of de Langhe, Fernbach, and Lichtenstein (2016, in this issue) to Simonson's (2015) assertions about the topic and suggests an agenda for future research. Follow-up research into the topic should emphasize the diversity of consumers and the multiplicity of their needs. It should recognize that reviews and ratings are complex social conversations embedded in consumers' multifaceted communicational repertoires. It should be cautious when using terms such as *objective* and *rational* when describing consumers and consumption. Being aware of the risks to external validity of studying average ratings may lead to frameworks with greater contextual integrity, and encourage collaborative communication between scholars from different perspectives working in this field.

Keywords: Amazon.com, contextual inquiry, netnography, online ratings, online reviews

Consumer-generated ratings and reviews play an important role in people's experiences of online search and shopping, influencing the consumption of movies, food, travel, automobiles, home services, technology, and much else. Laying out a broad course for "the BDT [behavioral decision theory] project," Simonson (2015, 32) proposes that researchers now focus on the interaction between consumer judgment and decision making and the "evolving information environment," which includes researching the use of these consumer-generated ratings and reviews. Simonson (2015) makes a range of broad assertions regarding their effects, including that they (1) have turned experience goods into search goods (21), (2) provide "rational" information about products' "harder" dimensions such as performance ratings and product use (23), (3) make "it easier for consumers to assess the quality (or "absolute value"; see Simonson and Rosen 2014) of products, potentially making brands less important" (22), and (4) "often represent the most reliable and most accessible predictors of product quality and user experience" (25). It is into this theoretical environment of assertions about the rational, reliable informational value of consumer-generated rating and reviews that de Langhe et al. (2016) position their study.

The purpose of this short article is not merely to applaud the thorough and thought-provoking response of de Langhe et al. to Simonson's assertions about the topic. In addition, I wish to respond and extend their article, showing how it

Robert V. Kozinets is the Jayne and Hans Hufschmid Professor of Strategic Public Relations and Business Communication at the University of Southern California's Annenberg School for Communication and Marshall School of Business; 3502 Watt Way, Los Angeles, CA, USA 90089; email: rkozinets@usc.edu. The author wishes to thank Eileen Fischer and Ulrike Gretzel for their important and helpful comments on earlier drafts of this article.

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

Advance Access publication XXXX XX, 2016

© The Author 2016. Published by Oxford University Press on behalf of Journal of Consumer Research, Inc.
All rights reserved. For permissions, please e-mail: journals.permissions@oup.com • Vol. 42 • 2016
DOI: 10.1093/jcr/ucv090

contains a range of underlying assumptions that might be tested, and setting out a succinct agenda for further research into the topic. First, this article emphasizes the multiplicity of consumer needs. Second, it points researchers to reviews and ratings' complex social communication environment. Third, it questions some of the assumptions of the de Langhe et al. study. Finally, it provides a framework to promote appreciation of different scholarly approaches to the topic and encourage collaborative communication between them.

APPRECIATING THE DIVERSITY OF CONSUMER NEEDS AND ONLINE RATING AND REVIEW USES

We must base our understanding of online ratings not on assumptions of their use but on knowledge of their actual real-world use by consumers. De Langhe et al. want to test Simonson's assertions about quality and "harder" dimensions, and so they choose to study Amazon ratings for "relatively vertically differentiated" product categories that "can be reliably ranked according to objective standards (e.g., electronics, appliances, power tools)" and for which "consumers typically care a lot about attributes that are objective in nature." Drawing on the power tool example, for instance, I conducted a quick netnographic (Kozinets 2002) scan of the DeWalt's CD970K-2 18-Volt Compact Drill/Driver Kit on Amazon, which was simply the first power tool to come up in my search. What I find is an immense wealth of information about many aspects of the power tool product experience.

The different types of questions people ask about the DeWalt drill reveal a panoply of power tool perspectives. The first question asks about the battery and its charger, the second about where it is manufactured, and the third about the case. Some people want to drill into concrete and plaster; others want to use certain types of drill bits. Some want to know if the batteries will power a flashlight, if the product will work "in Ireland," or if it comes with other options, such as a stud sensor. A variety of different needs and perspectives are represented. Hence even for something as apparently "vertically differentiated" as a power tool, there are many different aspects of the product itself, its packaging, case, lighting, electrical powering, service, warranty, prices, and brand image that might be salient to different people or even to the same person at different times or for different uses.

In unpublished ethnographic research that I conducted for a corporate client with consumers in their homes, I studied beauty product online shopping behaviors on sites that included Amazon.com. What I found was that shoppers carefully read textual reviews and examined specific, general, and average ratings. The more sophisticated, experienced, and motivated shoppers were interested in matching the product review and rating to a person who resembled themselves on some relevant dimension or dimensions—such as ethnicity, age, skin tone, eye color, hair color, or location (for seasonality concerns such as dry skin in winter). I observe the same sort of behavior on travel Web sites such as TripAdvisor. For example, people traveling with children of a certain age seek ratings and reviews of a hotel or destination from other people with children of the same or similar age. It would be difficult to classify this behavior as seeking an "objective" rating. On the contrary, it seems highly subjective.

As I develop later, this multiplicity of consumer needs complicates the notion that we can cleave "objective or technical aspects of product performance" from "more subjective aspects" (de Langhe et al. 2016). My beauty product shoppers were interested particularly in the performance and reliability aspects of products for their particular type of skin, face, or hair. Consumers want information that is very personal. Yet the performance characteristics they seek must be considered objective. Shoppers want to know the answer to a personal *and* performance-oriented question: "Will this product actually work *for me*, in my context? Will it do what I want it to do?" The multiplicity of consumer needs occurs because the realities of product use vary between individuals and contexts. Indeed, this is why we have segmentation, target, and positioning in marketing. Paying close attention to consumers' distinctive needs and characteristics, divergent perspectives, different subjective realities, and multifarious personal goals will help us conduct better research.

AMAZON RATINGS AS ELEMENTS OF A CULTURAL AND SOCIAL COMMUNICATION SYSTEM

Reviews and ratings offer consumers a social conversation, a communications environment that they use not only to talk about the objective and subjective characteristics of products and services, but also to socialize and communicate about themselves. Kumar and Benbasat (2006) find that providing customer reviews on Web sites improves customer perceptions of usefulness and social presence. Klaus (2013) studied the Amazon.com customer journey and found that for one customer, "Reading customer reviews is really helpful because it gives me more information about the book or product, but it is also interesting to know about the experiences of other people using the product" (448). Klaus (2013, 448) conceptualizes this aspect as "social presence," which "constitutes attributes reflecting the customer's virtual interaction with other shoppers through comments, product reviews, and social media linkages," and "was often cited with reference to its impact on the purchase decision process, in particular in the information search and alternatives evaluation stages" on Amazon.

Indeed, much of my early work studied and built theory about consumers' production and use of online reviews of media programs, food, clothing, and others products (Kozinets 1997, 1999, 2002). Although some of these products, like television shows and shots of espresso, might be termed "experience goods," consumers were nonetheless highly motivated to describe, assess, rate, teach, develop criteria, and demonstrate their evaluative expertise about their characteristics as well as to search out, share, and debate what were the highest quality offerings in the marketplace. The behavior of creating these reviews, sharing them, teaching one another about associated products and services, commenting, and complimenting them is both cultural (it communicates and bears meaning) as well as social (it creates affiliative connection between people). My concept of "virtual communities of consumption" (Kozinets 1999) has at its core the idea that online communications about consumption interweave pragmatic with social information: online "consumption knowledge is developed in concert with social relations" (254). Our continuing research must be constantly attuned to the social and cultural realities of consumer-generated online ratings and reviews, as well as the various social and cultural aspects of their creation, sharing, and use.

There is, in fact, a thriving and popular genre of humorous Amazon reviewing that demonstrates the rich complexity of the online rating and review environment. One of my favorite sets of reviews concerns a dedicated 1.5 meter cable made by Denon to connect DVDs and CD players to a Denon receiver. Amazon reviewers use their rating of the expensive cable humorously to extend their imaginations as well as to poke fun at people who take technical characteristics and the rating of technology products overly seriously. The following is an example that 4091 of 4144 people rated as helpful:

> "[Heading:] Great cable, but too fast. [Text:] Transmission of music data at rates faster than the speed of light seemed convenient, until I realized I was hearing the music before I actually wanted to play it. Apparently Denon forgot how accustomed most of us are to unidirectional time and the general laws of physics. I tried to get used to this effect, but hearing songs play before I even realized I was in the mood for them just really screwed up my preconceptions of choice and free will. I'm still having a major existential hangover." ("Frank Schulze" on Amazon.com)

Consumers have the option to contribute written text, questions, answers, comments, images, photographs, ratings of products, and ratings of review helpfulness on Amazon.com. In addition, they can use the Internet to search Facebook groups, YouTube videos, Twitter posts, podcasts, and forum discussions, as well as newspaper and magazine articles, Pinterest pins, other retails sites and their comments, or they can be brought there by hyperlinks, reviews, or other comments. They can use apps such as Yelp, Whatsapp, Instagram, and Snapchat to access reviews on the go, query friends and contacts in real time, and see photos of products in various contexts such as in other consumers' homes. Consumption of reviews and ratings occurs in this choice-saturated diverse communication context that Madianou and Miller (2012) call *polymedia*, "within which each individual medium is defined in relational terms in the context of all other media" (170). To be thorough and do justice to the entire and actual context in which judgment and decision making occur, future research into consumer-generated reviews and ratings must shift emphasis "from a focus on the qualities of each particular medium as a discrete context" to an understanding of it as part of an entire communication repertoire used by consumers attempting to balance constraints and goals with "the ways in which interpersonal relationships are enacted and experienced" (Madianou and Miller 2012, 170–71).

De Langhe et al. are correct in their conclusion that the Internet is not "making consumers more rational." The Internet is a remarkably complex and varied social and technological context that consumers use not only to share products' performance ratings and experiences, but also to engage, explore, connect, inform, and fulfill a wide range of social, communicative, emotional, and identity-focused needs. When people use Amazon.com's review and rating system, they use it socially. Although the software often is used as a source of peer opinion and information to inform decisions about potential purchases, it also acts as a platform for cultural connection, witty repartee, social commentary, entertainment, personal revelation, self-promotion, revenge seeking, and many other activities. Consumers' use of Amazon.com, like their use of every communication interface, is always a social and cultural experience. Thus it is vitally important that future consumer research recognize the complexity of the contemporary communications environment in which activities such as creating and using consumer reviews and ratings are embedded.

QUESTIONING THE ASSUMPTIONS OF ECONOMICS, OBJECTIVITY, AND SINGLE RATINGS

The term *objective quality* is rather slippery in the consumption context. Can we truly judge the absolute quality of a product like a vacuum cleaner in some objective and general sense that stands apart from individual consumers and their differentiated needs? For some vacuum consumers, being able to easily lift and carry the vacuum cleaner up and down three flights of stairs is of paramount importance. For others, it is the ability to handle large amounts of pet hair. Others want one that will store conveniently. Vacuum cleaners also have different design elements, colors, shapes, benefits, and sizes. Consumers have different physical sensitivities, brand preferences, and cultural tastes. Each different element of a product communicates meaning and offers value differently to different people.

Thus it seems that subjective perspectives and positions will always matter to consumers' evaluations of a given product's quality. There is no universal standard on which to base designations of real, true, or actual quality—and concomitantly no way to assess the so-called biases that allegedly detract from it. Although it may be useful for analytic purposes to create convenient fictions such as "search goods" and "vertically differentiated product categories" that exist in a dimension of undifferentiated consumer needs, we must be careful not to let our assumptions about how the world might be influence our ability to perceive how the world truly is. De Langhe et al. assume that some set of products can be largely functional, with their utility curves readily revealed, and then discover in a meticulous series of studies that (1) price efficiently predicts quality, (2) brands are matters of costs versus benefits, and (3) consumers want to be rational decision makers. These findings bear a remarkable resemblance to the assumptions of classical economics. However, they may not be valid as descriptions of a reality where a multiplicity of different consumer needs exist.

Further, I urge caution in posing and asking research questions regarding whether consumers use Amazon ratings—or any other reviews or ratings—"appropriately." The question contains within it the deductive whispers of a forthcoming right answer. The underlying assumption of the question is that reaching a decision based on objective ratings—as proxied by *Consumer Reports* scores—should be the goal of consumers' rating use. This assumption ignores or denigrates the many other uses to which people put ratings and reviews on Amazon.com, such as communicating with others, expressing oneself, and jokingly claiming that a cable can warp time and space. Further, it ignores the subjective filters through which people view all information.

To be realistic, future research into the topic should examine in realistic, naturalistic contexts how consumers create, share, and use ratings and reviews. How do consumers balance different types of information with the information on Amazon.com, *Consumer Reports*, and other sources? Are friends on Facebook viewed differently than experts on *Consumer Reports*? Do helpfulness ratings on Amazon.com matter? Are consumers suspicious of reviews that seem overly favorable or overly critical? What are the patterns in this performance? How do these practices vary—by expertise, age, gender, education, national orientation, online site and source? Further, research might inquire into the many nonmarket and nonrational uses of reviews, such as activism, advocacy, creative communication, and artistic self-expression.

It may be important for future researchers to appreciate that virtually all users of online ratings and reviews read and use text comments (Pavlou and Dimoka 2006). In one of the top-rated reviews of the DeWalt drill/driver kit I examined earlier, "Harv" (a pseudonym, following netnographic convention) warns people about the power tool's "smaller batteries," their nonreplaceable nature, its limited speed, and its workload limitations. However, Harv still gives the drill five stars, explaining in his review that he did so because "it finished the job with a lot of coaxing and still works despite melting something internally." Another top-rated review, by "Dave," waxes enthusiastic about the brand, tells us about the longevity of his last DeWalt drill, rates the new one "not a bad drill," then complains about the DeWalt Web site and online contact form as well as the product's batteries. Dave rates the drill one star out of five. My qualitative analysis of these reviews reveals an ongoing discrepancy between narrative reviews and the rather limited one to five star review format. Interpretation of ratings into reviews, and vice versa, by review creators and consumers is neither straightforward nor objective.

Studying the effect of the textual comments in a content analysis of over 10,000 of eBay's online auction marketplace reviews, Pavlou and Dimoka (2006) found that "buyers read and take into consideration feedback text comments to compensate for the inability of numerical ratings to offer detailed information" (408). Single-number ratings, it seems, are not particularly informative in and of themselves. Although it may seem efficient to boil down a variety of product attributes and their evaluations into a single number, this is a retailer practice that deserves careful scrutiny in our future research.

We must also question assumptions that ask us to reason from the general to the particular by taking the "average user rating" of a decontextualized "average consumer" and then assert that we can discover something relevant that might apply to the actual world of differentiated consumers with manifold needs. Instead, we should let the different needs, practices, perspectives, and experiences of our variegated world of consumers dictate the terms of our studies rather that any prefiguring presuppositions of objective desires, informational simplicity, or product utility. Developing the underpinnings of these presuppositions further in the next section, I deploy notions of ecological validity and contextual integrity to explore how to unite different approaches to researching online reviews and ratings.

EXTERNAL VALIDITY AND CONTEXTUAL INTEGRITY

In this concluding section, I wish to take a brief detour through the philosophy of science to discuss the tension between decontextualization and rigor that underpins much of the preceding sections. It is crucial to note, first, that science works through abstraction. Whether you conduct research as an economist, a psychologist, or an anthropologist, you take real-world events, such as people using Internet sites to buy products, and then abstract out from the multitude of potential variables a greatly reduced set to explore and explain. We reduce the complexity of the world so that we can study it. We turn reality into relationships between constructs; this is how theory is built.

Theories allow us to see the forest for the trees. In many cases, they allow us to see the various types of trees that compose the forest and, often, afford us insights into the relationship between them. No matter our approach, however, a certain risk of methodological reductionism always exists when we assume that we can understand the forest, or the phenomenon. A key issue is whether our mental maps of reality represent it well enough to be useful. Lynch (1982) points out that researchers must be especially careful about the external validity of theoretical tests when there are unmanipulated background variables that might interact with the manipulated independent ones. Simonson (2015, 29) contends that the current Internet environment presents us with an environment that connects many previously disconnected elements, such as the ratings and reviews of millions of other consumers. Thus what previously was noise may now be signal.

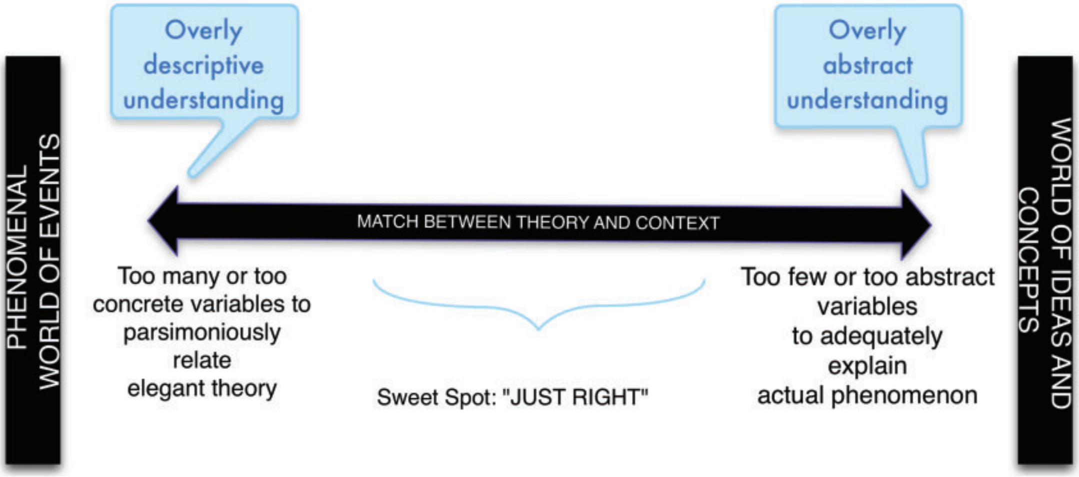
Favoring external validity over the internal variety, future research into this area should carefully examine actual occurrences of phenomena to guide the selection or creation of variables, data, constructs, and relations. Such an approach fits well with Simonson's (2015, 29) advice that "studies that excel on the external validity dimension should be allowed to meet lower (within a reasonable range) internal validity standards." It is nearly impossible to attain depth of understanding as well as representativeness, generalizability, internal validity, and the definitive ruling out of rival explanations in single experimental or cultural theory study. "But the field, as represented by journal editors and reviewers, may be better off being more lenient when evaluating studies that present potentially important findings, even if it is impossible to rule out some rival accounts" (Simonson 2015, 29).

When we engage with actual consumers, their needs, real marketing events, and sites of consumption and communication, researchers in ostensibly applied fields such as consumer and marketing research must struggle mightily with questions of external validity. In the service of rigorous method or parsimonious theorizing, we often abstract away considerable context. The question is, do we abstract so much of it away that we do violence to our ability to understand the actual phenomenon? I term this notion of balance between context and abstract theory *contextual integrity* and argue that the decision to theorize at a particular level is a very important, but largely unmentioned, aspect of social science discipline. To spur the conversation, Figure 1 illustrates the idea of a "Goldilocks approach" to matching theory with context. In that figure, we can see how adequate theorizing is a balancing act. Appropriate theory building occurs at a level of analysis poised between description and abstraction, between being overly complex and oversimplification, between being so close that it resembles a journalistic description and so distant that it is unrealistic and unrecognizable. The Goldilocks equilibrium lies in the middle and is, like the story's temperate bowl of porridge, just right.

How does figure 1 apply to the future study of online consumer-generated reviews and ratings? First it draws our attention to the appropriateness of the match between

FIGURE 1

FINDING THE RIGHT THEORETICAL BALANCE BETWEEN CONTEXT AND ABSTRACTION


PHENOMENAL
WORLD OF EVENTS
Overly descriptive understanding
Overly abstract understanding
MATCH BETWEEN THEORY AND CONTEXT
Too many or too concrete variables to parsimoniously relate elegant theory
Sweet Spot: "JUST RIGHT"
Too few or too abstract variables to adequately explain actual phenomenon
WORLD OF IDEAS AND CONCEPTS

theory, context, method, and analysis. In some cases, for some decisions, simple and elegant parsimonious measures and highly abstract theory-driven approaches may be sufficient for understanding. However, in the current ever-expanding universe of mobile and stationary Internet communications, which links everyone to everything, enabling two billion very different consumers to create, share, and consume a vast variety of social and emotional information, we face a far more complex consumption environment than ever before.

In this article, I have argued that the world of online reviews and ratings reveals customers with diverse needs and subject positions, whose narrative reviews may not agree with their numerical ratings. I have discussed consumers whose subjective guidelines determine the resemblance of particular reviewers to themselves so that they can determine from their review whether a particular product will actually work for them. In addition, I have shown how rankings and ratings on Amazon.com present consumers with a complex social communication environment, a conversation that grants them a sense of social presence as well as opportunities for a range of expressive and explanatory options. I have also noted how the review and ratings environment on Amazon, or any other site, is only one part of a wide range of options in contemporary consumers' current communicational repertoire.

Attention to appropriate fit between theory and context leads us to question the use of terms such as *objective*. It suggests that we not presume that one particular use of reviews or ratings is the correct or best one. It leads us to seek a deeper understanding of the multifaceted social, emotional, and relational properties of brands as they exist and expand in online reviews and ratings. It warns us to avoid using terms such as *rational* when describing consumers or their decision-making processes. Finally, it inspires us to initiate and value studies with high external validity, even if that means sacrificing some internal validity.

Simonson (2015) asserts, "the growing prevalence and impact of user reviews was not anticipated" by consumer researchers investigating the early years of the Internet. However, in 1999, I wrote that online consumers "create reviews of products, giving informed, justified 'thumbs up' or 'thumbs down' evaluations of [products and services and, by recognizing this,] marketers can have wide-ranging effects that inform and mediate consumer demand and consumption meanings across large numbers of others" (Kozinets 1999, 259–60). Many cultural and psychological consumer researchers such as Alladi Venkatesh, Russ Belk, Hope Schau, Janice Denegri-Knott, Mike Molesworth, Daiane Scaraboto, Marie-Agnes Parmentier, Nick Lee, Anne Schlosser, Wendy Moe, Donna Hoffman, Tom Novak, and Jonah Berger have been studying and developing a range of sophisticated contextualized theories to help explain how consumers create, share, and use online ratings and reviews. Our work should not be ignored or dismissed by BDT researchers, just as theirs should not be ignored or dismissed by us. De Langhe et al. perform a valuable service by following up on Simonson's (2015) call for more BDT work on the new informational environment, and by demonstrating how the Internet's effects on consumers cannot be easily explained as increasing their rationality or their resistance to branding. In the future, increased interaction and exchange of ideas between scholars of all disciplinary subfields working in this substantive area will undoubtedly be healthy for the continued growth and development of this burgeoning area of investigation.

REFERENCES

de Langhe, Bart, Philip M. Fernbach, and Donald R. Lichtenstein (2016), "Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings," *Journal of Consumer Research*.

Klaus, Philipp (2013), "The Case of Amazon.com: Towards a Conceptual Framework of Online Customer Service Experience (OCSE) Using the Emerging Consensus Technique (ECT)," *Journal of Services Marketing*, 27 (6), 443–57.

Kozinets, Robert V. (1997), "'I Want to Believe': A Netnography of the X-Philes' Subculture of Consumption," *Advances in Consumer Research*, Vol. 24, ed. Deborah J. Merrie Brucks and Deborah J. Merrie Brucks and MacInnis, Provo, UT: Association for Consumer Research, 470–75.

——. (1999), "E-Tribalized Marketing? The Strategic Implications of Virtual Communities of Consumption," *European Management Journal*, 17 (3), 252–64.

——. (2002), "The Field Behind the Screen: Using Netnography for Marketing Research in Online Communities," *Journal of Marketing Research*, 39 (February), 61–72.

Kumar, N. and Benbasat, I. (2006), "The Influence of Recommendations and Consumer Reviews on Evaluations of Websites," *Information Systems Research*, 17 (4), 425–39.

Lynch John G. Jr. (1982), "On the External Validity of Experiments in Consumer Research," *Journal of Consumer Research*, 9 (December), 225–39.

Madianou, Mirca and Daniel Miller (2012), "Polymedia: Towards a New Theory of Digital Media in Interpersonal Communication," *International Journal of Cultural Studies*, 16 (August), 169–87.

Pavlou, Paul A. and Angelika Dimoka (2006), "The Nature and Role of Feedback Text Comments in Online Marketplaces: Implications for Trust Building, Price Premiums, and Seller Differentiation," *Information Systems Research* 17 (4), 392–414.

Simonson, Itamar (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: What Really Influences Customers in the Age of (Nearly) Perfect Information*, New York: HarperCollins.

Imperfect Progress: An Objective Quality Assessment of the Role of User Reviews in Consumer Decision Making

ITAMAR SIMONSON

User reviews aggregate word of mouth and often greatly enhance consumers' ability to estimate product quality. Consumers decide for themselves whether and how to incorporate user reviews with other sources when evaluating options (e.g., a small minority uses *Consumer Reports*). Despite the extraordinary diligence of Bart de Langhe, Philip Fernbach, and Donald Lichtenstein and their use of a variety of data sources and methods, I have concerns about the purpose of the research, the evidence and distinctions they rely on, and the overstated conclusions. However, studying how user reviews and other currently available quality sources of information affect consumers is important and offers new directions for judgment and decision-making researchers.

Keywords: consumer reviews, consumer reports, perceived quality, internet, consumer decision making

There is no debate that user reviews have become influential in a growing number of product and service categories (Chevalier and Mayzlin 2006; Luca 2011; Ye, Law, and Gu 2009). While offline word of mouth (W-O-M) has had a good reputation as a source of information about quality (Herr, Kardes, and Kim 1991; Rosen 2009; Solomon 2013), online user reviews aggregate W-O-M of multiple buyers, primarily strangers, most of whom purchased and experienced the product being evaluated. It is noteworthy that, despite its long history, I am not aware of a study designed to determine once and for all whether W-O-M is a valid and reliable source of information about quality or whether W-O-M correlates with *Consumer Reports* (CR) ratings and resale prices. One wonders if that omission reflects the assumption that, despite the overall favorable reputation and unquestionable impact of W-O-M, its validity and reliability depend on many things, so it is not a meaningful research question.

Online reviews are highly accessible, often rich in detail and offered by knowledgeable consumers, and they are obtained at an exceptionally low cost. Beyond the star ratings, interested consumers can sort and select reviews and reviewers that address their specific concerns. However, according to the key conclusion of de Langhe, Fernbach, and Lichtenstein (2016, in this issue), user ratings tend to be unreliable and invalid and thus greatly overrated. As they write (Contribution Statement), "The broad conclusion of the article 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."

De Langhe et al. motivate the contribution of their research using a contrast with the message of a book I coauthored with Emanuel Rosen and subsequent related articles; as de Langhe et al. state, "User ratings allegedly provide an almost perfect indication of product quality with little search costs (Simonson 2015a, 2015b; Simonson and Rosen 2014a; 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 conclusions of de Langhe et al are based primarily on the low correlation between user ratings and CR ratings,

Itamar Simonson is Sebastian S. Kresge Professor of Marketing, Graduate School of Business, Stanford University (itamars@stanford.edu). This article benefited from the comments of Ran Kivetz and Emanuel Rosen.

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

Advance Access publication XXXX XX, 2016

© The Author 2016. Published by Oxford University Press on behalf of Journal of Consumer Research, Inc. All rights reserved.
For permissions, please e-mail: journals.permissions@oup.com
• Vol. 42 • 2016
DOI: 10.1093/jcr/ucv091

which they use as the measure of "objective quality," as well as on product resale values and certain statistical characteristics of user reviews (e.g., de Langhe et al. point out the limitations of small samples). The justification for using CR as the benchmark is that it is "the most widely used indicator of objective quality in the academic literature." Putting aside whether "objective product quality" is a meaningful concept (see later), CR's ratings have been used (also by me) as a convenient quality indicator, although CR has not previously been crowned as the "objective quality" authority. And the fact that the overwhelming majority of consumers ignore CR ratings while they often rely on user reviews suggests that, despite its long history, most consumers do not view CR as an important quality indicator.

Although certainly imperfect, product reviews of a sufficiently large sample of consumers can provide more relevant and reliable information that CR's technical tests may not be able to capture. Suppose, for example, that CR's experts rated a pair of headphones as having the best sound and a particular camera as having the best picture quality. However, according to 200 reviews of actual users of these products, the sound of the former and the picture quality of the latter are worse than some other options (evaluated by different consumers on the same Web site). Who would you rather rely on in making your purchase decisions, and should we infer from such inconsistency between CR ratings and user reviews that user reviews are invalid and unreliable? As discussed further later, it appears that most consumers, including CR subscribers, prefer to rely on users' reviews.

Despite their truly extraordinary diligence and persistence, I have serious concerns about the purpose of this research, the evidence de Langhe et al. rely on, and the overstated conclusions. The basic idea of testing whether user reviews are good or bad indicators of true quality and de Langhe et al.'s heavy reliance on the correlation between CR ratings and user reviews is, in my view, misguided. For one thing, although studying factors that moderate the value and impact of user ratings might be interesting, it is questionable whether any meaningful generalizations can be made about the overall value of user ratings.

Moreover, it is unclear why we should expect the within-rater evaluations of CR technicians, whose job it is to compare alternatives and rank-order them, to be highly correlated with the between-consumer experiences aggregated by user reviews (Hsee 1996; Nowlis and Simonson 1997). That is, the CR technicians will and probably should find differences along their chosen dimensions among the directly compared options even if actual (usually between-consumers) experiences associated with these products are rather similar across consumers and options. And where there are real significant differences, user reviews are quite capable and likely to show them, imperfectly of course, just as experiences are imperfect indicators of "objective quality."

A couple of clarifications are called for because, in the process of highlighting their contribution, de Langhe et al. did not accurately represent the position they are rebutting. It is true that Emanuel Rosen and I used the phrase "nearly perfect information" in our book title (2014a), *Absolute Value: What Really Influences Consumers in the Age of (Nearly) Perfect Information?* However, the book (see also Simonson and Rosen 2014b) makes it clear that the "nearly perfect information" term includes many other components beside user reviews, such as expert reviews, magazine reviews (CR being one of them), social media, YouTube, bloggers, and various other online sources. It is also inaccurate to state that Emanuel and I suggested that *user reviews* make consumers more rational and less reliant on brands, although we certainly discuss the implications of the current information environment with respect to what has been called irrationality or violations of value maximization. We also suggested that better, more accessible information sources about quality tend, in many but certainly not all categories, to diminish the impact of brands, past experience, prices, and other traditional quality cues. Briefly, the argument evolved from the conclusion that virtually all prior demonstrations of value maximization violations (termed "irrationality") reflected people's difficulty in handling absolute attribute values, leading them to a reliance on (relative) comparisons instead, which make consumers susceptible to various judgment and choice mistakes (Simonson 2008). Accordingly, to the extent that the current information environment makes it easier to evaluate the absolute values of individual options with less dependence on comparisons, consumers are arguably (1) less likely to "fall" for the classic irrationality demonstrations (e.g., context effects), (2) less dependent on often inferior quality cues such as brand names, and (3) on average, are likely to make better informed decisions.

Before examining more carefully the de Langhe et al. evidence, the "objective quality" and "vertically integrated categories" distinctions, and conclusions, it might be useful to put user reviews in (a somewhat speculative) historical perspective. In the beginning, perhaps around the time that humans appeared on earth and engaged in communication, people probably already shared (offline) W-O-M and offered recommendations regarding basic needs such as food, shelter, things to avoid, new tools, and un/desirable mates. Product names and prices probably followed soon after, and it is reasonable to assume that they were often used as quality and value proxies. Other quality cues, such as country of origin, probably emerged much later as trade and marketing evolved. Various new sources of information about quality were introduced in the 20th century, such as CR in 1936, JD Power and Associates in 1968, and *PC Magazine* in 1982. Then the Internet happened.

It is noteworthy that while early predictions about the impact of the Internet were on target in many respects (Alba et al. 1997), the growing prevalence and impact of user reviews was not anticipated, whereas other predicted trends such as a growing reliance on intelligent agents have (so far) turned out to be less significant than had been expected. As indicated, user reviews that aggregate W-O-M for most products and services and are highly accessible at a very low cost represent a wonderful development, even though they are certainly not perfect, in part because consumer experiences on which reviews are usually based are often ambiguous and possibly misdiagnosed. One limitation of Web user reviews has been a significant number of fake positive (e.g., by friends of the seller) and negative (e.g., friends of the owner of the restaurant across the street) reviews. However, as Emanuel Rosen and I (2014a, chap. 4) described, this problem has become less severe thanks largely to the efforts of the affected Web sites (e.g., Angie's List, Amazon, Yelp), which have been fighting this phenomenon, as well as the development of technologies to identify fake user reviews.

AN ASSESSMENT OF THE DISTINCTIONS AND EVIDENCE PERTAINING TO THE CORRELATION BETWEEN *CONSUMER REPORTS* AND USER RATINGS

De Langhe et al.'s main argument is not related to fake reviews but to their suggestion that user reviews are unreliable, invalid, and largely reflect the impact of brand names and prices. Fundamental to their argument is the distinction between products' "objective quality," presumably captured by CR, and "subjective quality," captured in user reviews. However, while there are dimensions for which the term *objective quality* trivially applies (e.g., a longer life light bulb is better), the distinction between "objective quality" and subjective quality of products is usually not meaningful or useful. CR and other such sources have no monopoly over "objective quality." The mere fact that the good people of CR or some other rating outfit decided to highlight certain features does not turn their criteria into "objective quality," unless consumers subjectively agree.

Moreover, the decision to designate CR as the "objective quality" judge is questionable, considering that CR's ratings are often inconsistent with the ratings of other "objective quality" contenders. For example, the most important product category for CR, which is the only one that has had its special publication, is cars. It turns out that CR car ratings are often remarkably uncorrelated with those of other "objective quality" authorities such as JD Power and Associates, and CR's methodology has come under severe criticism over the years (http://www.statesman.com/news/classifieds/cars/why-i-dont-rely-on-consumer-reports-reliability-su/nSxjT/; http://wheels.blogs.nytimes.com/2012/10/30/why-consumer-reports-and-j-d-power-are-so-different/; http://wot.motortrend.com/thread-of-the-day-do-you-consider-j-d-power-consumer-reports-ratings-before-buying-a-new-car-266719.html; http://latimesblogs.latimes.com/money_co/2009/03/edmundscom-vs-consumer-reports.html).

In fact, even with respect to de Langhe et al.'s flagship product example, baby car seats, CR's highly publicized ratings came under scrutiny, leading to one of CR's so-called scandals. As described on the Wikipedia page pertaining to CR,

> The February 2007 issue of *Consumer Reports* stated that only two of the child safety seats it tested for that issue passed the magazine's side impact tests. TheNational Highway Traffic Safety Administration, which subsequently retested the seats, found that all those seats passed the corresponding NHTSA [National Highway Traffic Safety Administration] tests at the speeds described in the magazine report. The *CR* article reported that the tests simulated the effects of collisions at 38.5 mph. However, the tests that were completed in fact simulated collisions at 70 mph. *CR* stated in a letter from its president Jim Guest to its subscribers that it would retest the seats. The article was removed from the *CR* website, and on January 18, 2007 the organization posted a note on its home page about the misleading tests. Subscribers were also sent a postcard apologizing for the error.

Although that particular mistake was unrepresentative of CR's ratings, it is consistent with the assumption that CR may have its own conscious or unconscious biases. For example, while I have no evidence for that, one might conjecture that in the face of increasingly intense competition from other sources as well as dwindling readership (http://jimromenesko.com/2013/10/30/change-is-hard-consumer-reports-restructures-to-survive-in-the-digital-era/), CR may seek opportunities to enhance its perceived value by highlighting product differences even when the distinctions have limited significance for actual consumer experiences. Thus even if (hypothetically) several baby car seats are all very safe, chances are that the CR technicians will still find justifiable ways to rate and rank-order the tested alternatives.

In addition to their distinction between objective versus subjective quality, de Langhe et al. state, "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)." The problem with this distinction, partially acknowledged (and then dismissed) by de Langhe et al., is that those vertically integrated categories tend also to have significant dimensions that are matters of taste. Putting aside obvious examples such as cameras and cars that have various taste-based features, even the flagship baby car seat example has some significant nonvertical dimensions. Having had some (slight to be precise) involvement recently in purchases of toddler car seats, I noticed that the perceived ease of getting the seat in and out of the car is an important consideration that is a matter of subjective perception and taste. More generally, it is quite rare to find purely vertical categories where CR can identify the best option for a diverse group of consumers.

Importantly, user reviews consist not just of ratings but typically also offer verbal evaluations. Some Web sites (e.g., Amazon) have systems to rate reviews and reviewers, and they order and highlight reviews accordingly. And many Web sites now enable consumers to sort and search reviews based on various useful criteria. The recent progress in analyzing textual data promises future improvements in consumers' ability to identify content more efficiently that suits their interests and priorities.

De Langhe et al. acknowledge the potential significance of textual analysis and narrative reviews but conclude that reviews' content tends to have limited impact compared with the more vivid, easy-to-process ratings. De Langhe et al. also cite anecdotally a couple of verbal reviews that were apparently biased and unrepresentative, and they mention that some consumers tend to pay more attention to negative reviews . However, the text and experience details can often be very helpful, with consumers deciding whether and how to process the additional information. Furthermore, the notion that "too much information" necessarily creates overload and can actually increase the impact of imperfect cues such as brand names mistakenly assumes that the quality assessment options available to consumers are either to process all reviews or none. As discussed in more detail elsewhere (Simonson 2015b; Simonson and Rosen 2014a), consumers have satisfactory intermediate solutions, and the available tools are making it easier to sort, refine, and use the selected portions of the available information quickly and efficiently.

One of the findings highlighted by de Langhe et al. is that user reviews are influenced by brand names and prices. It is unclear, however, why that finding is surprising or in what way it indicates that user reviews should not be relied on. For one thing, despite contrarian findings in certain categories (Mlodinow 2009), it is probably true that higher prices and more highly regarded brands tend to deliver, on average, higher quality experiences that are reflected in user ratings. Moreover, experiences are often ambiguous, and small "objective" differences do not affect perceived consumption experiences. Consequently, the higher expectations created by higher prices and highly regarded brands are indeed likely to affect perceived experiences, which are reflected in user reviews. Thus the finding that user reviews are more strongly influenced by brand names and prices than CR does not diminish their usefulness. To reiterate, user reviews are certainly not perfect indicators of "real quality," but with a sufficiently large number, they, on average, offer great value to consumers at a very low cost.

Finally, as previously reported (de Langhe, Fernbach, and Lichtenstein 2015), the same "biases" and "low reliability" of Amazon reviewers also apply to the ratings of CR's own subscribers on the CR Web site (i.e., product ratings by CR subscribers that are posted on ConsumerReports.com). This informative finding should have alerted the authors to the fact that consumers, whether on Amazon.com or on ConsumerReports.com, often form their own quality/experience assessments differently from the CR staff. In other words, even those who can easily access CR ratings when entering their own product ratings apparently disagree with CR and use different criteria than CR, making these consumers as susceptible to the claimed "biases" as those consumers who post reviews on Amazon. As an aside, the laboratory (Mechanical Turk) studies reported by de Langhe et al. (not discussed here) in an attempt to show that consumers adhere in principle to the CR criteria for assessing "objective quality" are not persuasive (e.g., due to demand effects) and cannot change the fact that consumers, including CR's own subscribers, often use different quality criteria and weights than CR.

HOW RELIABLE ARE RESALE VALUES AS THE BENCHMARK FOR EVALUATING USER REVIEWS?

Besides the CR benchmark, de Langhe et al. use correlations with resale prices posted on two Web sites, camelcamelcamel.com and usedprices.com (for the latter, only digital camera prices were examined), as further evidence that user ratings are not valid. It is difficult to assess what, if anything, we can learn based on the camelcamelcamel.com data because the authors provide very little detail (in the article and the online appendix) regarding how they used that source; for example, it is unclear how the sample of tested products and models was selected, how the numerous missing values were handled, and other important aspects. It is not possible to learn much from reported findings unless we know how they were obtained.

The authors provide a bit more detail about their use of the usedprice.com data source pertaining to digital cameras; they conclude that CR is more highly correlated with the camera resale values than user ratings. In principle, compared with CR ratings, resale values can have significant advantages because (1) they are more likely to reflect, at least to some degree, the market assessment of the product's consumption value, and (2) they are incentive compatible. However, there are various likely confounds that make it challenging to use resale values to compare the reliability of information sources.

Briefly, with the exception of CR, users and most magazine (e.g., *PC Magazine, Popular Photography*) reviews of new cameras tend to highlight the latest features offered by new models compared to previous and other recently introduced models. As we know, it does not take long for even better models to be introduced, at which time the previous new models are old news and dominated by the latest arrivals. Once the old models become relatively inferior, their resale values rapidly deteriorate, in all likelihood leading to a relatively low correlation between user ratings (entered primarily soon after product introduction) and resale values.

By contrast, CR evaluations of cameras are less focused on the latest feature advances. Moreover, they tend to be published long after the evaluated model was introduced, making these reviews less useful for consumers who are interested in getting the latest model. The review delay also means that CR ratings are likely to be published around the time that resale values are set. CR is more suitable for slower changing categories such as various appliances where the key attributes and features are rather stable, making the CR reviews potentially useful even if the recommended model was first introduced last year. Overall, there are various reasons to question whether correlations with available resale values allow us to evaluate the validity or reliability of user reviews.

CONCLUSIONS AND DIRECTIONS

Notwithstanding their thoroughness and hard work, I have two primary concerns about the de Langhe et al. research and conclusions: (1) I believe that the research question is misguided, and (2) the presented evidence rests on unsuitable distinctions and benchmarks. First, assuming a measure of true product quality were to exist, finding out whether this or that source of information is correlated with that selected measure is not a meaningful, *conceptually* interesting question. Even when (between the 1960s and 1980s) consumer researchers paid a great deal of attention to quality cues, such as price, brand, and country of origin, the focus was on how, when, and why consumers use these cues. The question of how accurate these cues were in predicting "true quality" or whether they were valid and reliable quality indicators did not receive much (or any) attention, perhaps because it is merely descriptive and there is no generally accepted way to measure real quality. Consumers get to decide what quality means and which sources they trust. For example, the fact that the overwhelming majority of consumers do not check CR ratings before making purchase decisions (and by all indications that percentage of consumers who consider CR is continuing to decline) suggests that they do not consider CR a particularly valuable source of information about quality.

Furthermore, it should have been obvious that the ratings of a technical team (or any group for that matter) comparing and then rank-ordering several options would be different from the ratings of individual consumers who typically consume and then evaluate just one option (i.e., they tend to be in the between-subjects condition). Prior research has already established that comparative tasks, not surprisingly, put more weight on comparable attributes, whereas separate evaluations emphasize "enriched" dimensions that can be evaluated on their own (Nowlis and Simonson 1997). We also know that experience tends to be ambiguous, and feature differences are often not felt or recognized when purchase decisions are made and/or during consumption (Nowlis and Simonson 1996; Thompson, Hamilton, and Rust 2005)—they may not make much difference, or even have a negative effect, in reality and are rated accordingly by consumers.

As indicated, there is also the challenge of defining and measuring product (and decision) quality. It seems unlikely that any objective, true, generally agreed-upon measure of product quality can be found in most cases, although some decisions (e.g., making an investment decision that is clearly dominated by another option) are unquestionably bad. De Langhe et al. argue that CR is the best representation of "objective quality" and rely on the fact that some earlier textbooks and articles used CR as an indicator of quality. Without repeating the previous discussion, I do not think that CR or the available resale values are adequate measures of true quality, so the degree to which other information sources, including user ratings, correlate with CR has limited relevance and diagnostic power. Having said that, I think that the de Langhe et al. article addresses an important area that should encompass not just user reviews but the broader changes in the consumer information environment.

User ratings, although certainly an imperfect measure, have some significant advantages (noted earlier) despite the "noise" and variability that affect individual and small samples of ratings. So it is not at all surprising that they have become so influential and are likely to become even more so (Luca 2011) as new advances make it easier to process textual inputs efficiently, identify capable and relevant reviewers, and screen out reviewers with ulterior motives. But as indicated upfront, the more interesting research questions are not about user ratings on their own but about the implications of the changing consumer information environment, an important component of which is user reviews.

I discussed elsewhere (Simonson 2015a, 2015b) a variety of specific research questions pertaining to the implications of the new information environment for consumer judgment and decision making. Without repeating or summarizing that discussion, a couple of questions alluded to by de Langhe et al. and by Lynch (2015; Schwarz 2015) deserve mention. Speaking of quality, a general question is whether the rich, new, and ever-expanding sources of information through the Internet and related communities allow consumers to make better decisions or, conversely, tend to have the opposite effect due to information and choice overload, echo chambers, and the like. Probably a more meaningful question focuses on interactions—studying factors that moderate and explain the conditions under which certain information conditions affect decision quality and decision making. But even that question uses decision quality as the dependent variable, which, as noted earlier, is inherently problematic and takes us back to the days when we focused on whether decisions were rational and value maximizing.

On the bright side, new research questions have emerged and can be studied based on the changing information environment and data sources. A key strength of the de Langhe et al. research program is the authors' ability to use diverse sources of data to test their proposition. More generally, currently available data types such as user reviews, social media (e.g., tweets, Likes, Followers), online search and purchase behavior, and consumer characteristics open a range of new research opportunities.

Over the past several years we have seen various investigations using user reviews and other information sources to explore a wide range of questions, including long-standing consumer behavior issues that can now be studied in a different context and using new types of data. Taking full advantage of these new opportunities may often require different skills and a greater collaboration with companies and researchers who can handle diverse types of data. Whether the changing environment, available data sources, and research tools also imply that the role of the classic experiment will start to diminish remains to be seen (and is beyond the scope of the current discussion). But data, tools, methods, and skills aside, the main challenge remains finding appropriate old and new research questions that deserve to be studied or revisited in the changing consumer environment.

REFERENCES

Alba, Joseph, Lynch, John, Weitz, Bart, Janiszewski, Chris, Lutz, Richard, Sawyer, Alan, and Wood Stacy (1997), "Interactive Home Shopping: Consumer, Retailer, and Manufacturer Incentives to Participate in Electronic Marketplaces," *Journal of Marketing*, 61, 38–53.

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

—— (2016), "Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings," *Journal of Consumer Research*.

de Langhe, Bart, Philip Fernbach, and Donald Lichtenstein (2015), "Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings," Leeds School, Colorado University, working paper.

Herr, Paul, Frank Kardes, and John Kim (1991), "Effects of Word-of-Mouth and Product-Attribute Information on Persuasion: An Accessibility-Diagnosticity Perspective," *Journal of Consumer Research*, 17 (4), 454–62.

Hsee, Christopher (1996), "The Evaluability Hypothesis: An Explanation for Preference Reversals Between Joint and Separate Evaluations of Alternatives," *Organizational Behavior and Human Decision Processes*, 67 (3), 247–57.

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.

Mlodinow, Leonard (2009), "A Hint of Hype, a Taste of Illusion," *Wall Street Journal*, November 20.

—— (1997), "Attribute–Task Compatibility as a Determinant of Consumer Preference Reversals," *Journal of Marketing Research*, 34 (May), 205–18.

Nowlis, Stephen and Simonson Itamar (1996), "The Impact of New Product Features on Brand Choice," *Journal of Marketing Research*, 33 (February), 36–46.

Rosen, Emanuel (2009), *The Anatomy of Buzz Revisited: Real-Life lessons in Word-of-Mouth Marketing*, New York: Crown Business.

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

—— (2015b), "The BDT Effect and Future: A Reply to John Lynch and Norbert Schwarz," *Journal of Marketing Behavior*, 1, 59–73.

—— (2014b), "What Marketers Misunderstand About Online Reviews," *Harvard Business Review* (January), 23–25.

Schwarz, Norbert (2015), "Which Mission? Thoughts About the Past and Future of BDT," *Journal of Marketing Behavior,* 1 (1), 53–58.

Simonson, Itamar (2008), "Will I Like a 'Medium' Pillow? Another Look at Constructed and Inherent Preferences," *Journal of Consumer Psychology*, 18, 155–69.

Simonson, Itamar and Rosen Emanuel (2014a), *Absolute Value: What Really Influences Customers in the Age of (Nearly) Perfect Information*, New York: HarperCollins.

Solomon, Michael (2013), *Consumer Behavior*, 10th ed., Pearson, 421.

Thompson, Debora, Rebecca Hamilton, and Roland Rust (2005), "Feature Fatigue: When Product Capabilities Become Too Much of a Good Thing," *Journal of Marketing Research* 42 (November), 431–42.

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

Ye, Qiang, Rob Law, and Bin Gu (2009), "The Impact of Online User Reviews on Hotel Room Sales," *International Journal of Hospitality Management*, 28 (1), 180–82.

Objective vs. Online Ratings: Are Low Correlations Unexpected and Does It Matter?

RUSSELL S. WINER
PETER S. FADER

The major point of the article by de Langhe, Fernbach, and Lichtenstein (2016, in this issue) is that objective ratings produced by *Consumer Reports* and online consumer ratings have a low correlation. We argue in this comment that this result is unsurprising due to some unresolved statistical issues, heterogeneity in terms of consumers' use of ratings and of the underlying consumer population and contexts, dynamics in the ratings system, and the complexity of modeling the generation of the consumer ratings. We also question why this low correlation matters given the fact that consumers use multiple sources of information, and more uncorrelated sources lead to more efficient decision making.

Keywords: online reviews, consumer reports

For over a decade now (Godes and Mayzlin 2004; Senecal and Nantel 2004), online reviews have provided an excellent opportunity for marketing and information scientists to better understand how customer sentiment impacts (and is impacted by) purchase behavior, marketing activities, and broader macro factors in the business environment. Among a large variety of topics, studies include how customer ratings affect demand (Chevalier and Mayzlin 2006 for books; Chintagunta, Gopinath, and Venkataraman 2010 for movies), how they are affected by other reviews (Moe and Schweidel 2012), and how the content of the review matters (Archak, Ghose, and Ipeirotis 2011). A count of the unduplicated citations of papers submitted to a recent research competition held by the Wharton Customer Analytics Initiative (WCAI) showed over 80 different empirical studies conducted using online reviews since 2004.

Online reviews are a subset of what is referred to as user-generated content (UGC) that also includes YouTube videos, blogs, and numerous other customer-created contributions to the online information environment. A number of UGC-based research papers—many featuring online reviews—were the focus of a special issue of *Marketing Science* based on a research competition sponsored by the Marketing Science Institute and the WCAI (Fader and Winer 2012).

Thus the article by de Langhe, Fernbach, and Lichtenstein (2016) is a welcome addition to the literature by taking a step back and attempting to analyze what the reviews actually mean to the consumers who use them. As the authors note, some recent work by Simonson (2015) and Simonson and Rosen (2014) has suggested that the power of this online information may be rising relative to the power of brands, or, in other words, the influence of marketing is waning relative to the influence of fellow consumers.

As a general point, this is not really a new idea. Many surveys in recent years examining the sources of information that consumers use for decision making put word of mouth, both online and offline, at the top of the list. For example, one survey found that 88% of consumers find online reviews "very influential" for purchasing new products

Russell S. Winer is the William H. Joyce Professor of Marketing at the Stern School of Business, New York University; Peter S. Fader is the Frances and Pei-Yuan Chia Professor of Marketing and the co-director of the Wharton Customer Analytics Initiative at the Wharton School of the University of Pennsylvania. We appreciate comments provided by Wendy Moe and David Schweidel.

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

Advance Access publication XXXX XX, 2016

© The Author 2016. Published by Oxford University Press on behalf of Journal of Consumer Research, Inc. All rights reserved.
For permissions, please e-mail: journals.permissions@oup.com • Vol. 42 • 2016
DOI: 10.1093/jcr/ucv092

from an unfamiliar brand and that 90% find Amazon reviews to be "determinative" in making purchase decisions (Lewis 2014). Another survey found that about 80% of Americans are influenced by online reviews (Bassig 2013). Not only do consumers look for objective information such as a restaurant's menu, but importantly, what other normally unknown consumers have to say about their experience with the product or service. De Langhe et al. ask two important questions: *What is the quality of this information,* and *should consumers be relying on it*?

Our interest in this comment is the first question. De Langhe et al.'s major finding is that the correlation is low between the online reviews and "objective" information as determined by *Consumer Reports* (CR). Our opinion is that (1) this finding is not surprising, and (2) that it does not really matter very much anyway.

THE DE LANGHE ET AL. FINDING IS NOT SURPRISING

What Is the Right Null Hypothesis Here?

Is 50–60% really a low degree of correspondence, as the authors strongly suggest? What, exactly, should we expect to see for the lines in Figure 4? Suppose we performed an analysis to see how well the user reviews "recover themselves." Specifically, let's use the observed distribution of reviews for each brand (which is readily available from Amazon but totally ignored by the authors). Let's sample values from the distribution for each pair of brands and see how often the resulting binary comparisons correspond to the observed averages. We suspect that the correspondence would not be very high—perhaps not at all different from the kinds of patterns shown in Figure 4.

Now let's think about the reverse null hypothesis analysis: how well would the CR ratings recover themselves? One obvious concern here is that there is no reported variation for them. This in itself is problematic: isn't there any measurement error in the way that CR determines its numbers, and shouldn't this be taken into account? The authors acknowledge that there could indeed be some unobserved variation in the CR data, but they dismiss this point by arguing that previous researchers have ignored it. That is a weak rationale, especially in this case where measurement error is far more central to the article than in the other cited cases.

But suppose we made reasonable assumptions about the measurement errors here and then repeated the same sampling exercise described earlier: How well would the sampled CR pairs recover the observed means? Again, we suspect that 60% might be on the high side.

Heterogeneity

One of the most well-documented findings from decades of analysis of consumer purchase data is the high degree of heterogeneity among consumers in terms of tastes and responses to marketing mix variables such as price. Of course, this is no surprise to consumer behavior scholars. Marketing scientists model this heterogeneity in two ways: observable heterogeneity, where it can be modeled as a function of measured variables, and unobserved heterogeneity, where it is as a result of some underlying stochastic process.

In the context of online reviews, observed heterogeneity can be a result of at least two factors.

Heterogeneity in the Use of Ratings Scales.

It is well known that people use ratings scales differently (Greenleaf 1992). The reasons for this could be yea-saying, bias, or other reasons. In this context, one consumer's 5 star rating is another's 3 star rating for equivalent "objective" quality. Normally, with multiple measures, one can correct for response-style differences with sufficient data within subject. Of course, although some reviewers become well known in their categories, generally, consumers looking at the reviews do not have such within-person data. In addition, in their analysis of 3 point scales, Lehmann and Hulbert (1974) show that even a 5 point scale may not capture all of the information that a reviewer wishes to convey. Many text-mining studies have shown, for example, that the valence in the text of a review has more information than the rating itself. De Langhe et al. acknowledge this as a limitation of their analyses. Clearly, given this well-documented behavioral phenomenon, the correlation between perceived and objective quality will diminish.

Heterogeneity of the Underlying Population and Contexts. It seems pretty clear to us that the degree of correlation between observed and rated quality will vary tremendously between different groups of consumers and different contexts. In a meta-analysis of studies of the effects of online reviews on sales elasticities, Floyd et al. (2014) found that critics' reviews, reviews on third-party sites, and review valence had the greatest impact on sales elasticities. There is every reason to believe that similar heterogeneity exists for online reviews. We expect that the correlation between objective and rated quality will vary depending whether it is an expert versus a consumer providing the ratings and on which site the reviews appear such as a third-party site like Amazon versus the brand's sites. For example, Schweidel and Moe (2014) found variance in brand sentiment for an enterprise software company across different social media. Whether the correlation also varies by the valence of the review is an interesting open question. Jang, Prasad, and Ratchford (2012) found that the influence of product reviews varies by the stage of the decision process. Anderson and Simester (2014) found that 5% of the reviews of the products of an apparel retailer lacked confirmed transactions and that these reviews were significantly lower than the reviews with confirmed transactions. Other kinds of heterogeneity such as cross-cultural and gender may also exist.

Dynamics. Another important area of marketing science research is in the area of dynamics. For example, many brand choice models incorporate updating mechanisms such as price expectations (Erdem and Keane 1996). An annual marketing dynamics conference focuses tightly on methods for (and implications arising from) the incorporation of time-varying factors in models of marketing phenomena.

Online reviews have also been found to be dynamic. For example, Godes and Silva (2012) decompose the changes over time in online reviews into two components: the length of time the product (in this case, a book) has been available for review and the pattern of prior reviews. The important conclusion from this work is that these two components are distinct and work differently to create what is generally a downward pattern for reviews. However, a particularly interesting finding is that once the researchers controlled for calendar time rather than the time the book has been available for review, the reviews actually became more positive, which is counter to normal observation.

The dynamics of reviews can affect the correlation between objective and rated quality in at least three ways. First, the correlation may be different early in a product's life cycle versus later. It is well known that early adopters are more knowledgeable than later adopters. As a result, early buyers may have different preferences than later buyers producing different reviews (Li and Hitt 2008). Thus the correlation between early adopters' ratings and objective quality may be higher than that of later adopters. Second, the process involving CR is itself dynamic. When a new product is introduced, it is rated by users before CR researchers test it in their lab. The information generated by CR is then available for later adopters. Thus we have a process of Ratings_t -> CR_{t+1} -> Ratings_{t+2}. Finally, the information set changes even without CR because buyers at any point in time have access to earlier ratings. For example, Kuksov and Xie (2010) developed a model of why and how firms might adapt their price and some aspects of the product offering in response to early reviews.

Complexity of the Modeling Process

De Langhe et al. develop a model attempting to explain the variation in user ratings as a function of the CR objective measures, price, and two dimensions of brand image—perceived benefits and brand affordability. All four variables were statistically significant with the non-CR measures explaining only 4.4% of the variance in ratings and the CR measures explaining only 1%. De Langhe et al.'s conclusion is that the combined effects of price and brand image are much larger than the effects of the CR scores.

Although we agree that better supported brands will likely have higher ratings, the model used to reach this conclusion is woefully underspecified and has significant statistical problems. Other authors have developed models to try to better understand how ratings are formed (Moe and Trusov 2011). The ratings formation process is undoubtedly highly complex and is a function of marketing variables such as advertising spending, prior ratings, consumer experience in the product category (i.e., expertise as we mentioned earlier), competition, and a number of other consumer-level, brand, and category variables. In addition, there are endogeneity issues with some of the independent variables (at least price, as de Langhe et al. point out) and a ratings decision-making process that implies a multiple-equation model (Hu, Koh, and Reddy 2014). Given the complexity of both the behavioral and modeling processes, why would we expect CR information to explain a significant amount of variance in the ratings?

DOES IT MATTER?

Perhaps the biggest question of all is whether it matters that the correlation between objective quality and rated quality is low. As we know, objective data are just one piece of information among many that some (but not all) consumers use to make decisions. In addition, don't we teach in core marketing classes that perceptions are what matter? Also, we do not believe that consumers are that naive to believe that each anonymous reviewer is somehow telling the "truth." At the same time, caveat emptor lives on.

In addition, in our view, it is good that the correlation is low. Two pieces of low correlated data are useful for decision-making purposes; if they are highly correlated, they are redundant, and using both will lead to inefficient behavior. The weights given each piece of information are determined by each consumer's "regression" based on his or her experience in the product category. In this vein, it would be interesting to test which predicts sales better, online ratings or CR ratings.

Interestingly, the low correlation between the ratings may be even more important to firms than to consumers. How should a marketing manager respond when the online reviews report lower ratings than objective measures made by the firm? The topic of online reputation management is becoming an important research area (Proserpio and Zervas 2015).

CONCLUSION

While we clearly take issue with some of the methods used (and inferences drawn) by de Langhe et al., we give them credit for the core idea that lies at the heart of their article. It is remarkable that no other researcher (to the best of our knowledge) has carefully investigated the interplay between objective CR evaluations and subjective UGC reviews. Further, we have no quarrel with the sampling methods used to create the comprehensive data set used in their article. In fact, we hope that de Langhe et al. will make this valuable resource available to other researchers because it could prove to be useful for a variety of research questions.

But in some sense, it is all a moot point: as we emphasized earlier, low correlations are not necessarily bad (besides being a fairly intuitive result in this case). When consumers "navigate by the stars," there is not necessarily a single path that will guide them to their goal. The ability to blend and balance CR ratings and UGC reviews can lead to greater decision flexibility and thus better outcomes in many cases.

REFERENCES

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

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

Bassig, Migs (2013), "Purchase Decisions of Close to 80 Percent of Americans Are Influenced by Online Reviews," www.reviewtrackers.com/purchase-decisions-close-80-percent-americans-influenced-online-reviews/.

Chevalier, Judith A. and Dina Mayzlin (2006), "The Effects 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.

de Langhe, Bart, Philip M. Fernbach, and Donald R. Lichtenstein (2016), "Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings," *Journal of Consumer Research*.

Erdem, Tülin and Michael Keane (1996), "Decision-Making Under Uncertainty: Capturing Dynamic Brand Choice Processes in Turbulent Consumer Goods Markets," *Marketing Science*, 15 (1), 1–21.

Fader, Peter S. and Russell S. Winer (2012), "Introduction to the Special Issue on the Emergence and Impact of User-Generated Content," *Marketing Science*, 31 (May-June), 369–71.

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 (2), 217–32.

Godes, David and Dina Mayzlin (2004), "Using Online Conversations to Study Word-of-Mouth Communications," *Marketing Science* 23 (Fall), 545–60.

Godes, David and José Silva (2012), "Sequential and Temporal Dynamics of Online Opinion," *Marketing Science*, 31 (May-June), 448–73.

Greenleaf, Eric A. (1992), "Improving Rating Scale Measures by Detecting and Correcting Bias Components in Some Response Styles," *Journal of Marketing Research*, 29 (May), 176–88.

Hu, Nan, Noi Sian Koh, and Srinivas K. Reddy (2014), "Ratings Lead You to the Product, Reviews Help You Clinch It? The Mediating Role of Online Review Sentiments on Product Sales," *Decision Support Systems*, 57 (January), 42–53.

Jang, Sungha, Ashutosh Prasad, and Brian T. Ratchford (2012), "How Consumers Use Product Reviews in the Purchase Decision Process," *Marketing Letters*, 23 (September), 825–38.

Kuksov, Dmitri and Ying Xie (2010), "Pricing, Frills, and Customer Ratings," *Marketing Science*, 29 (September-October), 944–57.

Lehmann, Donald R. and James M. Hulbert (1974), "Are Three-Point Scales Always Good Enough?" *Journal of Marketing Research*, 9 (November), 444–46.

Lewis, Truman (2014), Consumer Affairs web site, http://www.consumeraffairs.com/news/study-85-of-women-use-online-reviews-to-make-shopping-decisions-120914.html.

Li, Xinxin and Loren M. Hitt (2008), "Self-Selection and Information Role of Online Product Reviews," *Information Systems Research*, 19 (4), 456–74.

Moe, Wendy W. and David A. Schweidel (2012), "Online Product Opinions: Incidence, Evaluation, and Evolution," *Marketing Science*, 31 (May-June), 372–86.

Moe, Wendy W. and Michael Trusov (2011), "The Value of Social Dynamics in Online Product Ratings Forums," *Journal of Marketing Research*, 48 (June), 444–56.

Proserpio, Davide and Georgios Zervas (2015), "Online Reputation Management: Estimating the Impact of Management Responses on Consumer Reviews." Social Science Research Network (SSRN) download, http://dx.doi.org/10.2139/ssrn.2521190.

Schweidel, David A. and Wendy W. Moe (2014), "Listening in on Social Media: A Joint Model of Sentiment and Venue Format Choice," *Journal of Marketing Research*, 51 (August), 397–402.

Senecal, Sylvain and Jacques Nantel (2004), "The Influence of Online Product Recommendations on Consumers' Online Choices," *Journal of Retailing*, 80 (2), 159–69.

Simonson, Itamar (2015), "Mission (Largely) Accomplished: What's Next for Consumer BDT-JDM Researchers?" *Journal of Marketing Behavior*, 1, 7–32.

Star Wars: Response to Simonson, Winer/ Fader, and Kozinets

BART DE LANGHE
PHILIP M. FERNBACH
DONALD R. LICHTENSTEIN

In de Langhe, Fernbach, and Lichtenstein (2016), we argue that consumers trust average user ratings as indicators of objective product performance much more than they should. This simple idea has provoked passionate commentaries from eminent researchers across three subdisciplines of marketing: experimental consumer research, modeling, and qualitative consumer research. Simonson challenges the premise of our research, asking whether objective performance even matters. We think it does and explain why in our response. Winer and Fader argue that our results are neither insightful nor important. We believe that their reaction is due to a fundamental misunderstanding of our goals, and we show that their criticisms do not hold up to scrutiny. Finally, Kozinets points out how narrow a slice of consumer experience our article covers. We agree, and build on his observations to reflect on some big-picture issues about the nature of research and the interaction between the subdisciplines.

Keywords: online user ratings, perceived and objective quality, illusion of validity, statistical precision

The proliferation of user-generated content is the most important change to the consumer information environment in recent memory. As is apparent from the passionate responses to our work (de Langhe, Fernbach, and Lichtenstein 2016; hereafter, DFL), this development is of critical importance to all subdisciplines of marketing. At the same time, the field is being held back by a lack of crosstalk between the subdisciplines, an unfortunate state of affairs that we attribute to skepticism about one another's methods and misunderstandings about the goals of research. We are grateful to our editor Vicki Morwitz and to JCR for providing this outlet to debate the issues and for inviting some of the brightest lights from each of the subdisciplines to participate. We hope this is a spark that ignites more dialogue, argument, and collaboration within and across subdisciplines.

Our article conveys a simple idea: Consumers trust average user ratings as indicators of objective product performance much more than they should. As we have presented this work around the world, the response has run the gamut from intense interest and agreement to puzzlement to downright hostility and dismissiveness. The range of reactions is illustrated nicely by the commentaries. Simonson challenges the premise of our research. He raises a deep question about the nature of reality and consumer experience: If consumers want to optimize subjective experience does objective performance even matter? We think that it does and explain why in our response. Winer and Fader (2016; hereafter, WF) are also quite critical, arguing that our results are neither insightful nor important. We believe their reaction is due to a fundamental misunderstanding of

Forthcoming, Journal of Consumer Research Bart de Langhe, Leeds School of Business, University of Colorado at Boulder 419 UCB, Boulder, CO 80309, USA bart.delanghe@colorado.edu, Philip M. Fernbach, Leeds School of Business, University of Colorado at Boulder 419 UCB, Boulder, CO 80309, USA philip.fernbach@colorado.edu, Donald R. Lichtenstein, Leeds School of Business, University of Colorado at Boulder 419 UCB, Boulder, CO 80309, USA donald.lichtenstein@colorado.edu

All authors contributed equally and are listed in alphabetical order. We thank John Lynch, Gary McClelland, and Rick Netemeyer for comments.

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

Advance Access publication February 19, 2016

© The Author 2016. Published by Oxford University Press on behalf of Journal of Consumer Research, Inc. All rights reserved.
For permissions, please e-mail: journals.permissions@oup.com • Vol. 42 • 2016
DOI: 10.1093/jcr/ucw007

our goals, failing to appreciate the role of the consumer in our analysis. Our response focuses on dispelling their assertions and explaining why the results are not so easily dismissed. Finally, Kozinets provides a primarily positive reflection but points out how narrow a slice of consumer experience our article covers, and he suggests many future directions for research. We build on his observations to reflect on some big-picture issues about the nature of research questions and the interaction between the subdisciplines.

REALITY EXISTS AND CONSUMERS THINK SO TOO

Simonson raises many objections to our article, but we see a common thread running through most of them. He points out that consumers try to optimize subjective experience. User ratings, as direct measures of experience, should take precedence over scientific tests by experts if those tests do not match up with the ratings (assuming the average rating is relatively reliable from a statistical point of view). Thus our main message—that there is a disconnect between actual and perceived validity when it comes to objective performance—is immaterial. Kozinets raises a similar point when he asks, "Can we truly judge the absolute quality of a product . . . in some objective and general sense that stands apart from the individual consumers and their differentiated needs (Kozinets, 9)?" WF also raise this point, asking, "Don't we teach in core marketing classes that perceptions are what matter (WF, 9)?"

We are not surprised this issue came up in all the commentaries. It also comes up whenever we present the work, and we have grappled with it from the beginning of this research. In the article we acknowledge this point and tried to explain our perspective on it, but apparently more explanation is needed. The truth is we agree, to a point. Consumers care about subjective evaluations of the use experience, and these subjective evaluations may vary as a function of the product, the individual, and the context. As we say in the article, depending on a consumer's goals, she may want to focus on subjective evaluations over scientific tests (DFL, 14). However, Simonson takes the argument too far when he argues that it is not meaningful to distinguish between objective and subjective quality (Simonson, 6), and that consumers do not care about objective assessments of product performance (Simonson, 11). As we argue in the next section, it is beyond doubt that objective product performance can be measured and that consumers care about it.

The Age of (Nearly) Perfect Information?

Simonson provides a great example to illustrate the issue: imagine two hundred consumers rate a pair of headphones as having great sound quality, but *Consumer Reports* disagrees. Who should we trust? We don't have to imagine this. Consider Beats, the market share leader in high-end headphones, purchased by Apple for $3 billion in 2014. The Beats story is a phenomenal illustration of the power of traditional brand building. Beats allocates a lot of resources to marketing and celebrity endorsement, but they appear to cut corners when it comes to engineering. Hardware engineer Avery Louie conducted a teardown analysis of a pair of Beats headphones and found that the use of internal screws—which add production cost—was minimized in favor of less durable snaps and plastic fasteners (Louie 2015). More egregious, Beats appears to add nonfunctional, but heavy pieces of zinc to the headphones, presumably to fool consumers into thinking the construction is more solid than it is. While the headphones retail for $199, Louie estimates costs of good sold at just short of $17. The experts are not fooled. Scientific tests, including those conducted by *Consumer Reports*, rank Beats as mediocre in quality and a bad deal at such a premium price point (Eadicicco 2014).

Despite this, consumers love Beats headphones. The market share is tremendous. Ratings on Amazon are quite positive too. A search for all Beats over-ear headphones models with five or more ratings on Amazon.com reveals an average rating of four stars. This shouldn't be surprising. Consumers react to more than objective performance. They react to things like the emotional benefits they get from affiliating with celebrities, and the signaling value of wearing the coolest gear around. Most of them are not expert enough to truly evaluate the sound quality or to realize the heft that feels so good in the hand is due to useless chunks of metal. As Simonson points out, this is not exactly wrong. Their ratings reflect their experience. So, what's the problem? The answer is clear. In his own book, *Absolute Value: What Really Influences Customers in the Age of (Nearly) Perfect Information,* Simonson touts reviews as independent sources of information that make customers more informed, not suckers for clever marketing. Is this what the age of (nearly) perfect information looks like?

Here's another example. Kozinets (5) mentions his experience consulting in the beauty industry. He uses the example to motivate the idea that consumers look for information in reviews that is specific to their needs. Kozinets's point is that in some cases choosing between beauty products is a matter of taste, not performance. In those cases, there may be unique value in what other consumers have to say. However, the largest and fastest growing subsegment of the beauty industry by sales is skin care, responsible for about twice the sales of color cosmetics (Lopaciuk and Loboda 2013). Most of the growth in this subsegment is driven by functional products promising scientifically verifiable benefits like antiaging, wrinkle removal, and sun protection. Unfortunately, the skin care industry is a notorious cesspool of pseudoscientific jargon and unvalidated product claims. Beauty companies often tout their products as "clinically proven" despite no published clinical evidence, and they appeal to unproven biological and chemical mechanisms (Caulfield 2015). The industry is predicated on consumers' credulity. A perusal of ratings and reviews posted for antiaging creams on Amazon.com reveals the success of these marketing efforts. Ratings are consistently high, and many reviews parrot the dubious claims of the companies. A characteristic product, RegenFX Skincare Anti Aging Moisturizer Cream with Vitamin C, Vitamin E, Green Tea Extracts and Hyaluronic Acid, costs $42 for a 1-ounce vial and has an average rating of 4.6 stars. According to expert studies, including *Consumer Reports* testing (Consumer Reports 2011), the benefits of such products are similar to basic moisturizing creams that cost a tenth to a hundredth of the price.

Who really cares if people get worse audio performance or overpay for a tiny vial of skin cream, especially if they cannot even tell the difference? One constituency that cares is consumers. Consumers consult reviews to become more informed, not to be led to false conclusions about objective performance. Many of them want the best performance and would not like the idea of paying extra for an objectively inferior option, even if others enjoyed using it. These arguments take on even more weight in product categories where consumption choices have more serious consequences for welfare. Take, for instance, product categories that support health or safety. Be honest. Who do you want to trust when it comes to choosing car seats, bike helmets, sunblock, air filters, smoke alarms, or blood pressure monitors?

Another constituency that cares is policymakers. Consumer protection is predicated on the idea that happy consumers can still be injured. In a famous case, public policy officials were concerned that consumers believed the unsubstantiated claim that Listerine cures sore throats (Wilkie, McNeill, and Mazis 1984). We suspect if this controversy occurred today, many well-meaning consumers would be touting the sore-throat-fighting powers of Listerine in online reviews. That may be OK with Simonson, but it would be concerning to consumer protection advocates.

We Showed You Our Data, Now Show Us Yours

We have tried to stay out of the weeds by focusing on this one fundamental issue, but we conclude this section by considering some of the other criticisms in Simonson's commentary. Simonson makes some sweeping proclamations without the requisite data to back them up, a point also picked up on by Kozinets (2). Here are just a few examples. Simonson concludes that user reviews "often greatly enhance consumers' ability to estimate product quality" [abstract], that "online reviews are . . . offered by knowledgeable consumers" (Simonson, 2), that user reviews "offer great value to consumers at a very low cost" (Simonson, 9), and that "[*Consumer Reports*] may seek opportunities to enhance its perceived value by highlighting product differences even when the distinctions have limited significance for actual consumer experiences" (Simonson, 7). All of these assertions are proffered without a shred of evidence.

Simonson underestimates the technical capabilities and sophistication of *Consumer Reports*. We will not spend a lot of time defending them (they can do that themselves if they choose to). But it's worth noting that Simonson's casual dismissal of their capabilities reflects a disregard for huge swaths of the marketing literature that have used *Consumer Reports* as a benchmark on the basis of its validity, not just on the basis of precedent. His claim that "consumers . . . do not consider CR a particularly valuable source of information about quality" (Simonson 11) is nonsense. For instance, Tesla's stock price plummeted 6.6% the day after *Consumer Reports* withdrew its endorsement of the Model S sedan (Rogers 2015). In fact, Simonson inadvertently makes the point himself by highlighting an error in the Consumer Reports evaluations of car seats in 2007. Uri Simonsohn (2011) analyzed this very event in an article in the *Journal of Marketing Research* and found that consumer demand promptly responded to both the initial release and later retraction of *Consumer Reports'* evaluations, more evidence that consumers care about *Consumer Reports*.

Many, many criticisms of our methods and analyses are levied as if they are certainly true, without grappling with counterevidence and without considering the care with which we designed our studies. His challenge to our analysis of camera resale values is based on his intuitive model of camera obsolescence. Aside from not having any evidence for this counter-explanation (beyond his own intuitions), he also discounts the virtually identical results obtained using a different data set covering more than a hundred product categories.

Simonson also offhandedly dismisses all of our consumer studies as due to demand effects without any rationale or evidence for this claim. Demand effect criticisms are often leveled too easily (Shimp, Hyatt, and Snyder 1991). For a demand effect to drive a result, respondents must (1) detect some demand cue, (2) guess the hypotheses, and (3) decide to respond in compliance with the hypotheses. We don't see this as a plausible explanation for our results.

In our first consumer study, we simply asked participants to list reasons why they consult reviews and ratings across multiple product categories, without ever mentioning *Consumer Reports*. We then compared the information they provided with the dimensions covered by *Consumer Reports*. Respondents primarily listed objective quality dimensions, many of them covered by *Consumer Reports*. Where is the demand effect with this procedure?

The goal of consumer studies 2, 3, and 4 was to evaluate how strongly consumers use different cues such as price, average rating, and number of ratings to infer quality. In studies 2 and 3, participants went to real Amazon web pages, inspected products, and then judged the quality, in any way they wanted. In study 2, we asked consumers to predict *Consumer Reports* quality ratings. In study 3, we asked them to judge quality in general and also to judge purchase intention. In study 4, we orthogonally manipulated price, average rating, and sample size in a true experimental design, to rule out endogeneity issues. Across all three studies we found very similar results. Again, where is the demand effect that explains the consistent results across all of these studies?

It is unfortunate that Simonson so easily dismisses our consumer studies. The purpose of DFL is to compare the actual and perceived validity of average user ratings as measures of quality, a Brunswikian approach that has a long history in psychology and consumer research (Karelaia and Hogarth 2008; Lichtenstein and Burton 1989). Thus the consumer studies are absolutely critical to our arguments.

We are fully aware that no article can provide perfect or comprehensive data, and ours is no exception. But we did our best to present a range of data that provides converging evidence for our key ideas. That said, we are happy to be proven wrong. To Simonson we issue this challenge: show us the data.

TWO VIRTUES OF SIMPLICITY

WF make two criticisms of our article, that the findings are not surprising and that the results do not matter. Both criticisms are based on faulty assertions grounded in a fundamental misunderstanding of our research goals. Our goal is to compare the actual and perceived validity of average ratings as indicators of quality. To accomplish this goal, we analyzed many secondary data sources and conducted a series of consumer studies. Yet WF isolate and attack one piece of the evidence, the simple correlation between average ratings and *Consumer Reports* scores. Their biggest oversight, among many, is to ignore completely the critical role of the consumer in our analysis. WF's misrepresentation of our article has led to a confused and confusing litany of challenges that do not hold up to scrutiny.

WF's oversimplification of our evidence is ironic in that one of the major themes of their commentary is that our modeling is not complex enough. Our models do not specify a rating formation process, they do not account for consumer heterogeneity or dynamic changes in ratings over the product life cycle, and so on. We think that the simplicity of our analyses is a virtue, not a limitation. Isaac Newton wrote, "Truth is ever to be found in simplicity, and not in the multiplicity and confusion of things." We illustrate two senses in which Newton's words ring true in this case, and, in the process, demonstrate the flaws in WF's criticisms.

Virtue 1: Complexity Can Cause You to Lose the Forest for the Trees

The human mind has difficulty shifting between levels of analysis (Macrae and Lewis 2002). Thus one danger of complexity is that it can cause you to lose sight of the big picture. For instance, it's hard to think about the nitty gritty details of a model and simultaneously keep in mind the high-level structure of an argument or the conceptual coherence of a set of ideas. WF's speculation on "the right null hypothesis" (4–5) appears to be such a case. WF question how we should think about the correspondence level between Amazon ratings and *Consumer Reports* scores, "is 50–60% really a low degree of correspondence?"

We agree this is a critical question, which is why we dedicated so much of the article to addressing it. We have two important benchmarks in the article. The first benchmark is the most fundamental, consumer perceptions. Correspondence is low, not because it is low in an absolute sense, but because consumers believe it is much higher. There is a major disconnect between what average ratings actual convey and how consumers infer quality from them. This is a simple idea that is central to our message, but it is not considered by WF.

Price is another important benchmark. Price predicts *Consumer Reports* scores much better than does average user ratings. We find this result surprising because there is a substantial literature in marketing cautioning consumers about the weak relation between price and quality (often operationalized as *Consumer Reports* scores). The fact that average ratings are so much weaker seems important to us. In fact, consumers in studies 2, 3, and 4 trust average ratings much more than price, so they have the relationship reversed (DFL 12).

We find it perplexing that WF failed to consider both these benchmarks in their comments. The key results are summarized very early in the paper (DFL 2–3). Our best guess is that WF's focus on modeling details has caused them to miss the big picture. Rather than engaging with the benchmarks we provide, WF instead propose two new simulation analyses. From a mathematical perspective, WF do not appear to fully understand the analyses they are proposing or how they relate to analyses already in the article. Further, if we take a step back from the numerical details, it seems that WF do not appreciate how these analyses bear on our key claims.

Their first proposal is to analyze how well "reviews recover themselves" (WF 4) in the following way: take two products that each has a distribution of Amazon star ratings. Randomly sample one rating from each product and check which is higher. Repeat many times. Calculate the percentage of times the sampled rating is highest for the product with the higher average user rating. They "bet the correspondence would not be so high."

Although they do not refer to it in this way, the measure WF are proposing is called the "probability of superiority effect size" (Grissom 1994), or the "common language effect size" (McGraw and Wong 1992). Most consumer behavior researchers are probably more familiar with a measure of effect size called Cohen's d, which is computed by dividing the mean difference by the pooled standard deviation. Cohen's d is a linear transformation of the probability of superiority effect size. Both Cohen's d and probability of superiority are also directly related to the area under a receiver operating characteristic (ROC curve) (Ruscio and Mullen 2012), a measure of classification accuracy that may be more familiar to the marketing science community.

WF asked us to simulate this measure, but it is not necessary to do a simulation to compute distribution overlap. The combinatorics of a 5 point scale are straightforward, and the percentage superiority can be determined simply, as follows: $[\#(x > y) + .5\#(x = y)] / n_x n_y$, where # is the count function and x and y are vectors of scores for the two products. Or, even more simply, they could have just asked us to compute the average Cohen's d. We computed probability of superiority for all within-category pairwise comparisons of products, and the correlation with Cohen's d was 0.94. The reason the correlation is not exactly 1 is because Amazon ratings are not normally distributed, but, for all intents and purposes, WF are asking us to compute the average Cohen's d and use this as a benchmark to assess the correspondence between average user ratings and *Consumer Reports* scores. They seem to believe that this will provide a novel perspective on our results, but this does not make sense.

WF's confusion is indicated by their claim that we "totally ignored" the distribution of ratings in our analyses (WF 4). This is a surprisingly blatant mischaracterization. Analysis of the ratings distributions is a centerpiece of the article. For instance, we analyze how correspondence between average user ratings and *Consumer Reports* scores changes as a function of standard error of the rating distribution (DFL 6), and in a follow-up analysis, as a function of sample size and standard deviation (DFL 6). In another important analysis presented in the General Discussion (DFL 13), we look at how often pairwise *t* tests between average user ratings for two randomly chosen products are significant. Our analyses show that correspondence is lower when standard error is higher (DFL 6), that *t* tests are not significant about half the time (DFL, figure 4, 13), and that correspondence is related to the significance of the *t* tests (DFL, figure 4, 13).

All of these analyses are intimately related to effect size. The major difference is that our analyses also take into account the role of sample size in addition to the averages of the distributions and their standard deviations. (For instance, the *t* statistic is Cohen's d divided by the square root of the sample size). The results indicate that effect sizes are often too small relative to sample sizes to conclude much. Yet consumers happily jump to strong quality judgments regardless of the sufficiency of the sample sizes, as we show in consumer studies 2, 3, and 4 (DFL 12). This is one of the main reasons that consumers overestimate the validity of average ratings.

We are now in position to consider how WF's proposed analysis bears on our key argument, and we reach an ironic conclusion: WF are absolutely right that the average effect size is small, as is apparent from analyses already in the article. But they fail to appreciate that this *supports* our key claim that consumers overestimate the validity of average ratings. In fact, it is a central pillar of our argument.

The second benchmark proposed by WF is to examine how well *Consumer Reports* scores would recover themselves using a similar simulation, given reasonable assumptions about error in *Consumer Reports'* measurements. They "suspect that 60% might be on the high side" (WF, 5). Although this benchmark is conceptually more meaningful than average effect size, their 60% claim is way off. Suppose that the true quality score of a product lies within 10 points of the score determined by *Consumer Reports* with uniform probability. This would be a huge measurement error, given that the median range of *Consumer Reports* scores across product categories in our data set is 31. A simple simulation reveals that the ranking of the scores posted by *Consumer Reports* would converge with the ranking of the true quality scores 79% of the time. A recovery rate of 60% would imply true quality scores that lie within 35 points (!) of the scores posted by *Consumer Reports,* greater than the *range* of scores of most categories.

Moreover, *Consumer Reports* rates products on multiple dimensions and then averages these subscores to arrive at a composite quality score. In the article we show, via simulation, that random variation to the weights *Consumer Reports* assigns to the sub-dimensions has little effect on the composite score (DFL 8). An analogous argument applies to error in measuring the sub-dimensions. If varying the weights of the sub-dimensions while holding constant the scores has little effect on the composite measure, then adding measurement error to the scores while holding constant the weights should also have little effect. If $e_x = (x/\mathrm{b}) * e_b$ then $\mathrm{b} * (x + e_x) = (\mathrm{b} + e_b) * x$, where e_x is the random component added to the subscore x and e_b is the random component added to the weight b. Thus, without any additional data, a careful reading of the analyses in the article shows that WF's criticism based on measurement error in *Consumer Reports* scores is severely overstated.

WF's substantive purpose for suggesting these analyses is to argue that our results are not surprising. Surprisingness is a notoriously slippery concept. What one person finds obvious may be astonishing to another (Lynch 1998). WF suggest that the surprisingness of our results should be judged against the intuitions of marketing science scholars. We disagree. Our goal is to understand whether consumers have correct intuitions about the validity of online ratings, so we are much more interested in what they think.

Virtue 2: Simplicity Permits Empirical Generalization

Models can serve various functions. In consumer research, models are usually aimed at supporting empirical generalization by identifying factors that explain behavior and are invariant across contexts. WF point out many things our models do not do (e.g., model the rating formation process, capture dynamics and heterogeneity, etc.). They see this as a problem, but we see it as a necessity. The goal of the article is to compare the actual and perceived validity of average user ratings as measures of quality, so we modeled factors that consumers may use when making quality inferences. Most consumers have no way of assessing heterogeneity, dynamics, or the review formation process when consulting online ratings. They tend to use simple choice processes. This is another example where WF have failed to consider whether their criticisms actually speak against our key claims. Taking the perspective of the consumer, it is clear that many of the issues that WF perceive as limitations of our research only make our key points stronger. Not only is the average rating a poor predictor of quality overall, but its usefulness depends on a host of contextual factors that most consumers have no way of evaluating.

One benefit of simplicity is that simple models often work well in the real world (Dawes 1979). Complex models can overfit data and perform poorly when used to predict out-of-sample observations. For example, Wübben and Wangenheim (2008) compared the relatively complex retention model of Fader, Hardie, and Lee (2005) that models heterogeneity in customer retention to a much simpler "hiatus" model by fitting data sets from multiple industries. The simple model performed better than or equal to the complex model in all cases. Our goal is not to impugn Fader et al.'s model, which we admire and teach in our customer analytics course. Our point is that complexity and generalization do not always play nicely together.

Brighton and Gigerenzer (2015) refer to the preference for complex models as the "bias bias" because faith in complex models often reflects neglecting the variance component of the bias-variance tradeoff. Fortunately, there seems to be increasing awareness of these issues in empirical studies, particularly those analyzing big data. We have seen several presentations recently where most of the focus is on "letting the data speak for themselves" through basic summary statistics and model-free evidence. We applaud these developments.

DFL is inspired by a simple but compelling idea called the "illusion of validity" (Tversky and Kahneman 1974). An illusion of validity occurs when one overestimates the predictive value of a cue because the cue seems representative of the outcome of interest. One reason we were so drawn to this topic is because we feel this illusion ourselves, even now. We see an average rating that we know is flawed but still want to trust it. That is the crux of the article. It is a simple idea that deserves simple treatment.

One of the developers of this idea, Amos Tversky, sometimes remarked that he was not a very sophisticated mathematician. His colleagues and students found this claim laughable because he was the best applied mathematician that any of them knew (Steven Sloman, personal communication, December 2015). Tversky's talent was not in mathematical complexity. It was in simple ideas expressed as simple models that explain behavior across a wide range of contexts. A first-year undergraduate would have no problem following the math behind prospect theory (Kahneman and Tversky 1979), support theory (Tversky and Koehler 1994), or the contrast model of similarity (Tversky 1977). We are not comparing our work to Tversky's (anyone making that comparison would come out sorely lacking). The point is that there is a huge difference between simple and simplistic.

Do the Results Matter?

WF conclude by questioning whether our results matter. They argue that the low correspondence is a feature, not a bug, because consumers now have two uncorrelated sources of quality information to inform their decisions. The problem with this argument is that consumers do not aggregate information in this way. As is clear from our consumer studies, they go to ratings primarily as a free proxy for the kind of information provided by *Consumer Reports*, and they think they are getting it. Moreover, they jump to unwarranted conclusions based on insufficient sample sizes. Again, the problem is not the low correspondence; rather, it is the disconnect between what consumers think they are getting and what they are actually getting.

Here's another way that these results matter. Our understanding of the new information environment has major implications for how companies should allocate resources. The results that WF so easily dismiss—the positive influence of high prices and strong brands on ratings, and the low correspondence of ratings to objective quality indicators like *Consumer Reports* scores and resale prices—suggest that companies should not be so hasty to shift resources away from traditional marketing and branding, as suggested by recent articles in influential outlets like *Harvard Business Review*, *The Economist*, and *The Wall Street Journal*. Importance is another concept in the eye of the beholder, but it strikes us that businesses might be interested in a better understanding of the antecedents of ratings.

WHERE IS THE GOLDILOCKS ZONE?

By necessity, the tone of this commentary has been confrontational so far. Taking the lead from Kozinets, we will attempt to elevate the discussion in this final section. While Kozinets clearly takes issue with some of our claims, we appreciate that he also attempts to be positive in the sense of offering new data and insights to support his claims (e.g., the netnography of power tools, his experiences consulting for beauty products) and suggesting directions for future research. We agree with his overarching theme. Our article only covers a narrow slice of the consumer experience. Although the average star rating is an important driver of consumer behavior, Kozinets rightly points out that reviews serve many other purposes. He is also right that consumers look for information that is specific to their own needs, and such information cannot be gleaned from the overall average. These points should spur new research ideas. How do consumers navigate and integrate all these different pieces of information? The answers to these questions can fill many dissertations, and we hope they will.

Kozinets goes on to discuss the philosophy of science and offers a useful figure depicting an arrow that spans from the highly descriptive "phenomenal world of events" to the highly abstract "world of ideas and concepts." This distinction is closely related to the trade-off between complexity and generalization we discussed earlier. The more complexity you put into your model, the more descriptive it is of a particular context and the less it captures abstract concepts that are invariant across contexts. He suggests that researchers should try to stay in the middle of the arrow, in the "Goldilocks zone" that strikes the right balance between complexity and generalization.

This reminds us of the ending of the SpongeBob SquarePants movie (yes, two of us have toddlers). Viewers have been led to believe that SpongeBob's home, Bikini Bottom, is a good size town. As the perspective shifts to the world of humans, the camera pans out, and we see that all of Bikini Bottom is contained in about a square meter of ocean. We are not comparing any marketing scholars to sea creatures. The point is that the world looks a lot different to the denizens of Bikini Bottom than it does to the people standing on the beach. Similarly, we all live in different places on Kozinets's arrow. To each of us, our little neighborhood feels much bigger and more comprehensive than it is. What to one of us feels like highly descriptive research may seem hopelessly abstract and disconnected from reality to someone with a different orientation.

The idea of a Goldilocks zone contains within it the whispers of a directive. We are not sure that researchers should be in the business of telling other researchers what questions to ask and the methods they should be using to address them. It strikes us as futile to try to define a single level of analysis that we will all agree constitutes a Goldilocks zone. It's also probably counterproductive. It is fairly easy to argue that research along the entire extent of the arrow has value if done competently. On the abstract side this is obvious; consider Einstein imagining himself riding on a light wave. The other side of the arrow is more contentious, but many people find value in highly descriptive approaches, for instance in the work of phenomenologists like Husserl and Heidegger. As Kozinets points out, due to the pervasive role that user-generated content plays in the lives of consumers nowadays, the issues are so multidimensional and complex that many types of research are needed to understand them.

These ideas are especially important to keep in mind in an interdisciplinary field like marketing. The topic of online reviews and ratings clearly has interdisciplinary appeal, which is a good thing. But interdisciplinarity also introduces a risk of imposing one's favorite constructs and methodologies on others' work (Shugan 2002). We have to be careful not to evaluate research in terms of whether the theory and methods used in an article fit with our mental model of what an article should be like. Instead we should be asking whether the approach is appropriate to address the specific research question the researchers are asking. Obviously there is also an onus on researchers to be clear about what they are trying to accomplish. Keeping these points in mind may help us build a more cumulative and integrative science.

REFERENCES

Brighton, Henry and Gerd Gigerenzer (2015), "The Bias Bias," *Journal of Business Research*, 68 (8), 1772–84.

Caulfield, Timothy (2015), *Is Gwyneth Paltrow Wrong About Everything?: How the Famous Sell Us Elixirs of Health, Beauty & Happiness*, Boston: Beacon Press, 2015.

Consumer Reports (2011), "Wrinkle Creams: Miracle or Mirage?" http://www.consumerreports.org/cro/magazine-archive/2011/september/health/wrinkle-creams/overview/index.htm.

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

de Langhe, Bart, Philip M. Fernbach, and Donald R. Lichtenstein (2016), "Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings," *Journal of Consumer Research*, 42 (6), doi:10.1093/jcr/ucv047.

Eadicicco, Lisa (2014), "Apple Just Paid $3 Billion for a Company That Makes Really Mediocre Headphones," http://www.businessinsider.com/beats-headphones-quality-2014-5.

Fader, Peter S., Bruce G. S. Hardie, and Ka Lok Lee (2005), "Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model," *Marketing Science*, 24 (2), 275–84.

Grissom, Robert J. (1994), "Probability of the Superior Outcome of One Treatment over Another," *Journal of Applied Psychology*, 79 (2), 314–16.

Kahneman, Daniel and Amos Tversky (1979), "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, 47 (2), 263–92.

Karelaia, Natalia and Robin M. Hogarth (2008), "Determinants of Linear Judgment: A Meta-analysis of Lens Model Studies," *Psychological Bulletin*, 134 (3), 404–26.

Kozinets, Robert V. (2016), "Amazonian Forests and Trees: Multiplicity and Objectivity in Studies of Online Consumer-Generated Ratings and Reviews," *Journal of Consumer Research*, doi:10.1093/jcr/ucv090.

Lichtenstein, Donald R. and Scot Burton (1989), "The Relationship Between Perceived and Objective Price-Quality," *Journal of Marketing Research*, 16 (February): 429–43.

Lopaciuk, Aleksandra and Miroslaw Loboda (2013), "Global Beauty Industry Trends in the 21st Century," paper presented at the Management, Knowledge and Learning International Conference, Zadar, Croatia.

Louie, Avery (2015), "How It's Made Series: Beats by Dre," http://blog.bolt.io/how-it-s-made-series-beats-by-dre-154aae384b36#.shn3uqye2.

Lynch, John G. (1998), "Presidential Address: Reviewing," *Advances in Consumer Research*, 25, 1, 1–6.

Macrae, C. Neil and Helen L. Lewis (2002), "Do I Know You? Processing Orientation and Face Recognition," *Psychological Science*, 13 (2), 194–96.

McGraw, Kenneth O. and S.P. Wong (1992), "A Common Language Effect Size Statistic," *Psychological Bulletin*, 111 (2), 361–65.

Rogers, Christina (2015), "*Consumer Reports* Pulls Recommendation on Tesla Model S," http://www.wsj.com/articles/consumer-reports-pulls-its-recommendation-on-the-tesla-model-s-1445363667.

Ruscio, John and Tara Mullen (2012), "Confidence Intervals for the Probability of Superiority Effect Size Measure and the Area Under a Receiver Operating Characteristic Curve," *Multivariate Behavioral Research*, 47 (2), 201–23.

Shimp, Terence A., Eva M. Hyatt, and David J. Snyder (1991), "A Critical Appraisal of Demand Artifacts in Consumer Research," *Journal of Consumer Research*, 18 (3), 273–83.

Shugan, Steven M. (2002), "The Mission of *Marketing Science*," *Marketing Science*, 21 (1), 1–13.

Simonsohn, Uri (2011), "Lessons from an "Oops" at *Consumer Reports:* Consumers Follow Experts and Ignore Invalid Information," *Journal of Marketing Research*, 48 (1), 1–12.

Simonson, Itamar (2016), "Imperfect Progress: An Objective, Quality Assessment of the Role of User Reviews in Consumer Decision Making," *Journal of Consumer Research*, doi:10.1093/jcr/ucv091.

Tversky, Amos (1977), "Features of Similarity," *Psychological Review*, 84 (4), 327–52.

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

Tversky, Amos and Derek J. Koehler (1994), "Support Theory: A Nonextensional Representation of Subjective Probability," *Psychological Review*, 101 (4), 547–67.

Wilkie, William L., Dennis L. McNeill, and Michael B. Mazis (1984), "Marketing's 'Scarlet Letter': The Theory and Practice of Corrective Advertising," *Journal of Marketing*, 48 (2), 11–31.

Winer, Russell S. and Peter S. Fader (2016), "Comment on 'Navigating by the Stars'," *Journal of Consumer Research*, X (X), XX–XX.

Wübben, Markus and Florian v., Wangenheim (2008), "Instant Customer Base Analysis: Managerial Heuristics Often 'Get It Right'," *Journal of Marketing*, 72 (3), 82–93.