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

 • 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, 836)?" WF also raise this point, asking, "Don't we teach in core marketing classes that perceptions are what matter (WF, 848)?"

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, 830). However, Simonson takes the argument too far when he argues that it is not meaningful to distinguish between objective and subjective quality (Simonson, 842), and that consumers do not care about objective assessments of product performance (Simonson, 844). 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 (835) 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 (834). Here are just a few examples. Simonson concludes that user reviews "often greatly enhance consumers' ability to estimate product quality" [abstract 840], that "online reviews are . . . offered by knowledgeable consumers" (Simonson, 840), that user reviews "offer great value to consumers at a very low cost" (Simonson, 843), 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, 842). 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 844) 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" (847) 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 828).

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 818–819). 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 847) 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 847). 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 822), and in a follow-up analysis, as a function of sample size and standard deviation (DFL 822). In another important analysis presented in the General Discussion (DFL 829), 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 822), that t tests are not significant about half the time (DFL, figure 4, 829), and that correspondence is related to the significance of the t tests (DFL, figure 4, 829).

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 828). 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, 847). 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 824). 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 $e_x = (x/b) * e_b$ then $b * (x + e_x) = (b + e_b) * x$, 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.