How Do People Use Star Rating Distributions?

Jingqi Yu
jingqi.yu@rotman.utoronto.ca
Rotman School of Management
105 St George St, Toronto, ON M5S 3E6 Canada

David Landy
dhlandy@gmail.com
Netflix
Los Gatos, California, USA

Robert Goldstone
rgoldsto@indiana.edu
Department of Psychological and Brain Sciences
1101 E 10th St, Bloomington, IN 47405 USA

Abstract

It may seem pointless to compare two products with the exact same average rating and total number of reviews without other review information. Now imagine a scenario in which the distribution of star ratings is also available to decision makers in addition to these two attributes. Will the decision still be uncertain as it is before or the distributions of stars will engender a preference towards one of the products? To answer this question, the current study used variability of star ratings as an approximation of a product's distribution. The behavioral studies showed that participants exhibited distinctive choice patterns when the distribution of ratings was provided even when the average rating and total number of reviews were the same between two products involved in a comparison. A utility-based cognitive model was therefore developed to identify the underlying mechanism as to why people chose the way they did.

Keywords: rating distributions, average rating, total number of reviews, variability, modeling

Introduction

The proliferation of user-generated content (UGC), and electronic word-of-mouth, has reshaped the way in which people communicate with information and with each other. One way in which UGC has greatly influenced people's lives is by providing consumers with online reviews on which they can rely in order to make product choices. As such, product ratings have increasingly become a major factor in how people buy products. While we know much about how consumers incorporate average ratings (valence) and number of reviews (volume) -- two of the most frequently examined and arguably used dimensions -- into their choice processes (Powell et al., 2017; Rosario et al., 2016; Floyd et al., 2014; Chevalier & Mayzlin, 2006), we do not know nearly as much about how they evaluate riskiness of ratings for individual products. In everyday consumer contexts, it is not uncommon for consumers to encounter situations in which they have to choose between one product that receives ratings that are consistently moderately good and another related product, whose ratings are usually excellent but occasionally terrible. To illustrate this idea, imagine the following scenario: you are presented with Product A and Product B, both with an average rating of 3 stars (out of 5 stars) based on 400 reviews. Asking you to pick one product out of these two may perplex you as they have identical average ratings and number of reviews. You have every right to be indifferent in this case. However, what if now more information about the two products to you: Product A's 3-star rating is averaged across 400 3-star individual ratings while Product B's 3-star rating is averaged across 200 1-star individual ratings and 200 5-star ratings. With this new information, do you have a preference towards one of the products? This new preference you may feel illustrates a key tenet of this research -- there is more than meets the eye when it comes to online reviews. Meanwhile, it also highlights the informational role of the third "v", in addition to valence and volume, in online reviews – variance.

Variance as Social Cue: Opinion Dispersion

While the notion of variance may sound more statistical than social, this dimension of information has been investigated under the guise of opinion dispersion. Before online reviews became ubiquitous, many researchers have examined how traditional word-of-mouth (WOM) influences choice behavior and how this process is affected by other distributional characteristics. WOM seems to be one of those buzzwords that is all around us in this Web 2.0 era. In marketing and communication literature specifically, WOM is defined as "oral, person-to-person communication between a receiver and communicator whom the receiver perceives as non-commercial, concerning a brand, a product or a service" (Arndt, 1967). Through the lens of WOM, rating variance reflects the degree of social consensus, or lack thereof, in experiences of the product. Consensus is then captured by the level of agreement, or homogeneity, among opinions: People agree that the product is good, or bad, or mediocre.

In a perfect world, each product receives homogeneous reviews from fellow consumers and the true quality can thus be easily recovered. Yet, in reality, oftentimes products end up with highly divergent ratings such that average ratings can be misleading. *Mulan* (Caro, 2020) falls under the latter category as the 33122 reviews (at the time of writing) it has received on Google so far are clearly distributed in a bimodal fashion with spikes at 1-star and 5-star ratings, resulting in an average rating of 2.7 stars (see Figure 1).

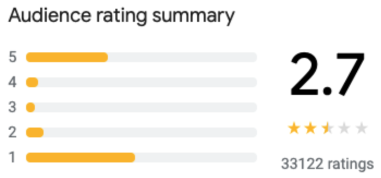


Figure 1. Rating distribution of Mulan (2020) at the time of writing.

Apparently, some audiences love the movie while others do not. At first glance, 2.7 is a mediocre, if not poor, rating. If one interprets this score as representative of product performance, s/he is likely to be misled: ratings of 3 are actually the least common. This phenomenon is not unique to

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Mulan (2020). As pointed out by Amendola et al. (2015), movies are becoming increasingly controversial, and thus bimodal. Maybe there is also the question of whether social consensus exists among products with bimodal distributions (e.g., J- or U-shaped), which in fact are encountered rather frequently "in the wild" (Hu et al., 2006; 2009).

When people consider the experiences of others regarding a product they may purchase online, what they are really doing is setting an expectation for themselves -- what would constitute a satisfactory versus an unsatisfactory outcome and what will the product be like for *me*? West and Broniarczyk (1998) examined how consumers integrated critic opinions (differed in levels of consensus) into their product evaluations They reported that consumer response to critic consensus depends on product quality relative to individual expectations. Specifically, when the average critic rating was below consumers' expectation disagreement (i.e., high variance among critic ratings) was preferred. They reason that this is because disagreement indicates that there is still a possibility (albeit small) of exceeding the expectation. When critics agreed with each other and their consensus was below the consumers' aspiration level, however, the chance of achieving a desired level of satisfaction is asymptotically zero. To sum up, when the average critic rating is below consumers' expectation, preference for disagreement is driven by the desire of potentially getting a really pleasant experience. On the other hand, when the average critic rating was above consumers' expectation, consensus (i.e., low variance among critic ratings) was preferred. They reason that this preference is primarily driven by consumers' desire to not fall below their expectation, even if it means they would probably miss the alternative that delivers the best experience. This is just one of the many examples reflecting the idea that the influence of opinion dispersion can be affected by summary statistics such as valence and volume (Khare et al., 2011; He & Bond, 2015). To use the language of the prospect theory (Kahneman & Tversky, 1979), the rest of the paper also refers to choosing a low-variance alternative as being risk-averse and choosing a high-variance alternative as being risk-seeking.

Overview of Studies

This research takes a different tack by framing online decision-making with consumer ratings as risky choices and studying it with a combination of modeling and behavioral experimentation. Risky choice behavior is by no means a new topic. A large body of literature has documented how decision makers behave when facing risks and/or uncertainties (see Fischhoff & Broomell, 2020 for a review). However, comparatively, we know much less about risky choice behavior in terms of online consumer choice. While a small number of previous studies has examined the role of rating variance in consumer decisions, very few studies have employed a modeling perspective with individual experimental data. Anecdotal evidence has also highlighted the possibility that consumers may not use ratings in a similar fashion. Hence, instead of answering the question of whether rating variance impacts product evaluation, this research seeks to identify whether and in what ways consumers differ in how they use the standard five-star rating scale. Through exploring the utility of individual star ratings, this research provides a more fine-grained picture of how people evaluate rating variance by showing that different uses and interpretations of the rating scale could lead to different risky choice behavior in online consumer decision-making. The variance of star ratings is used as a proxy measure of riskiness; the more variable ratings are for an individual product, the riskier that option is. As such, preferring more-variable products is interpreted as demonstrating risk-seeking behavior and preferring less-variable products as risk-averse behavior. More broadly, this research contributes to the literature on decision-making by incorporating risky choice behavior in online consumer contexts with an emphasis on how decision makers evaluate the riskiness of a product.

When we think of how people use and interpret the rating scale, one intuitive way is that people would focus on extreme ratings; that is, they would pay more attention to the number of 1-star and 5-star ratings. This intuition has been supported by anecdotal evidence that people are apt to focus more on these two rating categories than the 2-, 3-, and 4-star ratings. Therefore, one plausible interpretation of the rating scale is that people would assign a large psychological distance between 1- and 2-star ratings as well as 4- and 5-star ratings. This would yield the pattern of 1---2,3,4---5. For people whose interpretation largely follows this pattern, they are expected to show risky choice behavior, depending on average ratings. Specifically, when the average rating is low, they are expected to be more risk-averse, hoping to minimize the number of 1-star ratings. When the average rating is high, however, they are expected to be more risk-seeking, hoping to maximize the number of 5-star ratings. The opposite pattern is also highly possible, due to reasons such as the proliferation of fake reviews. Because 1-star and 5-star ratings are often the source of fake reviews (Luca & Zervas, 2016), it is possible that people decide to just discount them. In this case, they would compress the utility differences between 1- and 2-star ratings, as well as 4- and 5-star ratings. For these people, their purchase decisions are not solely driven by one type of star-rating category. Instead, they are expected to prefer a product that has a smaller number of 1- and 2-star ratings altogether or a greater number of 4- and 5-star ratings altogether.

The above two possibilities related to situations when people perceive the 1- and 5-star ratings to be symmetric in the sense that they either expand or compress the utility differences between these extreme ratings and 2-, 3-, and 4-star ratings (aka., middle ratings). There is also the possibility that people pay attention to only one of the extreme star-rating categories. For example, people could perceive 1-star ratings to be uniquely bad, whereas 4-star ratings are just as desirable as 5-star ratings. When the average rating is low, the purchase decisions of these people are expected to be primarily driven by the total number of 4- and 5-star ratings; when the average rating is low, their decisions are expected to be primarily affected by the number of 1-star ratings. In contrast, people could also perceive 5-star ratings to be uniquely good, whereas 2-star ratings are just as undesirable as 1-star ratings. For people who categorize star ratings in this way, they would primarily base their decisions on the total number of 1- and 2-star ratings when the average rating is low, and the number of 5-star ratings when the average rating is high. In addition to people's perception regarding the relationships between extreme and middle ratings, another common difference that might distinguish people is how they categorize 3-star ratings. While some people may perceive it to be more negative than positive (thus more similar to 2-star ratings), others may perceive it to be more positive than negative (thus more similar to 4-star ratings).

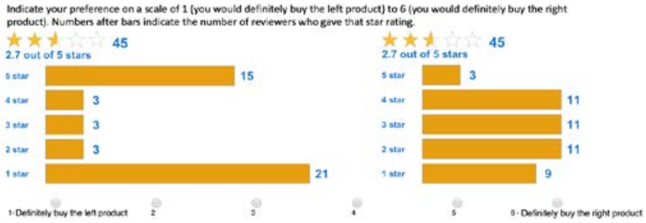
Facing these various possibilities of rating scale interpretations, this research employs a combination of computational cognitive modeling and behavioral experiments to identify some of the common ways people ascribe meanings to the standard five-star rating scale. A principal component analysis (PCA) was applied to obtain a synthetic picture of potential criteria people apply to categorize star ratings. Model performance was reported from both a quantitative (likelihood ratio testing) and qualitative (explanation of observed human data) lens.

Experiment

The present experiment used the standard five-star rating system (description format) to examine the influence of average ratings, number of reviews, and rating variance on consumers' product preference.

Participants. We recruited 211 undergraduates at Indiana University Bloomington in exchange for course credit.

Materials. The experiment was a within-subject experiment manipulating three *v*s of online reviews: valence, volume, and variance. Valence featured four levels: extremely low (1.X), low (2.X), medium (3.X), and high (4.X). Both volume and variance featured three levels – low, medium, high – with each having its own unique rating profile. Each product pair shared the same valence and volume and differed only in their variance. This rule led to 36 comparisons. Ratios between counts of rating categories were maintained with increased volume (e.g., 20 out of 45 were scaled up to 60 out of 135). Three sets of comparisons of these kinds were created in such a way that each level of every attribute took three possible values (e.g., a low valence: 2.6, 2.7, 2.8), amounting to 108 trials. A product's rating profile resembled that of Amazon.com, displaying the three *v*s. The lengths of horizontal bars were created proportionally to reflect relative frequencies of rating categories.



Indicate your preference on a scale of 1 (you would definitely buy the left product) to 6 (you would definitely buy the right product). Numbers after bars indicate the number of reviewers who gave that star rating.

Figure 2. A sample trial with a valence of 2.7 and a volume of 45. The left product's bimodal ratings are more variable than the right product's unimodal ratings.

Procedure. The experiment consisted of 81 randomized and counterbalanced trials. Participants were tasked to indicate their purchase preference on a 6-point Likert scale, ranging from *1* = definitely buy the left product to *6* = definitely buy the right product (see Figure 2 for a sample trial).

Results

Behavioral results. We refer to a product with a higher (lower) rating variance as a more (less) - variable product. Responses were first recoded as binary decisions between the less-variable (1) and more-variable (0) products. This treatment allowed us to measure the probability of picking a less-variable product. For simplicity, this paper uses this probability as the reference probability, therefore denoting it as *P*. At an aggregated level, choice of rating variance seemed to be by chance, $P = .05$, $\chi^2(1) = 1.45$, $p = .23$. Now we ask if this aggregated analysis masked any stable individual differences. Figure 3 shows that the distribution of experimental responses is wider than that of a binomial distribution with a probability equal to the average probability of choosing a less-variable product observed in the experiment ($P = .50$).

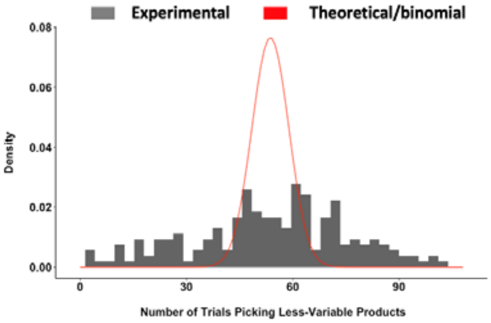


Figure 3. The comparison of the experimental and a binomial distribution (P = .50,).

A chi-square goodness-of-fit test confirmed that this binomial distribution is indeed not a good fit for the experimental data ($p < .001$). Therefore, there were stable individual differences in preferences, rather than just randomness around an overall group preference, namely choosing options with smaller variance as a result of risk aversion. This suggests the importance of examining mental models of rating systems that give rise to individual differences in preference as opposed to concluding a general preference for products with lower rating variance.

Cognitive Model

The behavioral results were only informative to the extent that we know participants indeed differed in their preferences when rating distributions were available. However, it is unclear as to why they differed. The goal of this section, therefore, was to introduce a utility-based cognitive model to reveal underlying mechanisms of this observed difference.

Model Development For the purpose of this exploration, the relative difference between any two adjacent star ratings was more important than their absolute values (the smaller the difference, the more similar two star-ratings were perceived by an individual user). Consequently, the current model's free parameters represented utility differences between any two adjacent star ratings. The model was constrained on both the lower and upper ends where the utilities of a 1-star and a 5-star rating were fixed to 1 and 21, respectively. Hence, the present model featured three free parameters, with each corresponding to the utility difference between any two adjacent star ratings. Although $\mathit{\Delta}_{45}$was also fit, it was not counted as a free parameter because it could be calculated with 20 - $(\mathit{\Delta}_{12} + \mathit{\Delta}_{23} + \mathit{\Delta}_{34})$. As the present model used utility theory and prospect theory as its theoretical backdrop, it is referred to as a utility-based model for simplicity. (Prospect-theory-style probability rescaling was attempted but did not improve model fit.) The model is mathematically presented below:

$$U \ = \sum_{i=1}^{5} p_i \times u_i$$

where

$$u_i = \begin{cases} 1 & i = 1 \\ 21 & i = 5 \\ u_{i-1} + \Delta_{(i-1)i} & i = 2,3,4 \end{cases} \quad \text{and} \quad p_i = \frac{c_i}{\sum_j^5 c_j},$$

$\mathit{\Delta}_{12}, \mathit{\Delta}_{23}, \mathit{\Delta}_{34}, \mathit{\Delta}_{45} \geq 0$

The rule choice rule: $P_i = \frac{{U_i}^{\gamma}}{\sum_{j=1}^{2} {U_j}^{\gamma}}$

$\gamma = 40$ (uncertainty parameter)

$\Delta_{(i-1)i:}$ Subjective utility difference between star-rating $(i-1)$ and i

u: Subjective utility of a star-rating

U: Subjective utility of a product

c: The count of a star-rating

p: The probability of a star-rating

P_i: The probability of selecting product I when given the two choices

Principal Component Analysis (PCA). To identify dimensions along which subjects differed in terms of how they interpreted the five-star rating scale, a Principal Component Analysis (PCA) was conducted. Table1 presents the loading of the four PCs. Overall, judging from the results of PCA, our initial intuition was confirmed such that people tended to differ in how they view the utility differences between extreme and middle ratings, as well as how they position 3-star rating relative to 2- and 4-star ratings.

Table 1

Loadings of Four PCs

	PC1	**PC2**	**PC3**	**PC4**
Δ_{12}	-0.23	0.74	0.63	<.001
Δ_{23}	0.59	-0.62	0.51	<.001
Δ_{34}	-0.96	-0.20	-0.19	<.001
Δ_{45}	0.79	0.43	-0.43	<.001

In the following sections, we illustrate how our cognitive model is capable of explaining and predicting risky choice behavior. In the present paper, choosing a less-variable product is considered as exhibiting risk-averse behavior and choosing a more-variable product is considered as exhibiting risk-seeking behavior. Principal component analysis (PCA) identified three principal dimensions based on estimated parameters: Dim. 1: $\mathit{\Delta}_{12}, \mathit{\Delta}_{34}$ vs. $\mathit{\Delta}_{23}, \mathit{\Delta}_{45}$ (middle clumps: whether 3s were grouped with 2s or 4s); Dim. 2: Δ_{12}, Δ_{45} vs. Δ_{23}, Δ_{34} (edge expanders vs. middle expanders); Dim. 3: Δ_{12}, Δ_{23} vs. Δ_{34}, Δ_{45} (edge expanders: differences in which half of the rating scale was expanded). Each dimension separated the five star-categories into three groups. The results are first summarized in Table 2 and then three representative cases (one from each dimension) were presented to graphically explain model prediction.

Table 2

Relative positioning of the five star-categories.

Recurring Dimension	One End	The Other End
Middle clump	1,2 —— 3,4 —— 5	1——2,3—— 4,5
Edge expanders vs. middle expanders	1,2 —— 3 —— 4,5	1——2,3,4—— 5
Half expanders	1,2,3—— 4 —— 5	1——2——3,4,5

Notes. Numbers that are connected by commas indicate star categories that are perceived relatively similar in terms of their utilities. An em dash (——) indicates greater utility differences. For example, 1,2 —— 3,4 —— 5 represents the usage in which people compress the utility differences between 1- and 2-stars and 3- and 4-stars but expand the differences between 2- and 3-stars as well as 4- and 5-stars.

Table 2 summarizes the overall patterns in relative positioning of the five star-categories. As described above, these diverse usages of the standard five-star rating scale would make different predictions. When the average rating is extremely low (1.X), 1- and 2-stars are the main constituents. When 1-stars are singled out (Column 3), decision makers would generally show risk-averse behavior because they want to avoid 1-stars as much as possible due to their significantly low utilities, even compared to 2-stars. This inclination leads them to choose a product with more stars in the middle over a product with more stars on the two ends. On the contrary, for decision makers who perceive 1- and 2-stars to be approximately equally bad (e.g., Column 2), they are more likely to prefer products with fewer 1- and 2-stars all together. In our experiment, more-variable products bore this characteristic when average ratings were extremely low, and therefore choosing them led to risk-seeking behavior. When the average rating is high (4.X), 4- and 5- stars are the main constituents, making their relative positioning more prominent. For decision makers who perceive 5-stars significantly more valuable than 4-stars, that is, singled out 5-stars (e.g., Row 3, Column 3), they would be more likely to prefer a product with more 5-stars, despite having more 1-stars. This tendency produces risk-seeking behavior. In contrast, for those who perceive 4- and 5-stars to have roughly equal values, they are more likely to prefer products with more 4- and 5- stars all together. In our study, because less-variable products bore this characteristic when average ratings were high, choosing them led to risk-averse behavior. These trends are illustrated in Figure 4.

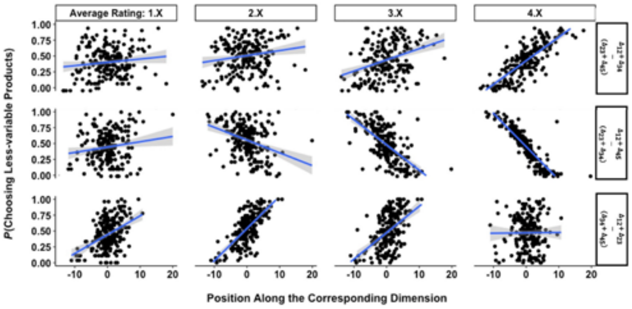


Figure 4. Values along the y-axis represent each individual's selection probability for less-variable products. Each point at every panel represents one individual subject. The x-axis aims to capture where a subject was located along each principal dimension. For Dim. 1, the x value was calculated with (Δ_{12}+ Δ_{34})- (Δ_{23}+ Δ_{45}); (for Dim. 2, the x value was calculated by (Δ_{12}+ Δ_{45})- (Δ_{23}+ Δ_{34}); for Dim. 3, the x value was calculated by (Δ_{12}+ Δ_{23})- (Δ_{34}+ Δ_{45}). The closer a y-value is closer to 1, the more risk-averse one's choice appears to be, the closer a x-value is to 0, the more risk-seeking one's choice appears to be. The horizontal facet labels (1.X, 2.X, 3.X, 4.X) represent different valence levels, or average ratings. The vertical facet labels (Dim1, Dim2, and Dim3) represent three PCs.

Dim 1. (Middle clump) Explanation: This dimension determined which end-star-category, namely 1-stars and 5-stars, was on its own. Individuals on one end of this dimension ($\Delta_{12} + \Delta_{34} < \Delta_{23} + \Delta_{45}$) assigned relatively lower values to Δ_{12} and Δ_{34}, and relatively higher values to Δ_{23} and Δ_{45}, exhibiting a 1,2 — 3,4 — 5 pattern (Figure 4). This notation represents rating compression with utility distances where numbers that are connected by commas indicate star categories that are perceived relatively similar in terms of their utilities. An em dash (—) indicates greater utility differences. Hence, people with a 1,2 — 3,4 — 5 tended to compress the differences between 1- and 2- stars as well as 3- and 4-stars but expand the differences between 2- and 3-stars and 4- and 5-stars. To their mind, 1- and 2-stars were bad, 3- and 4-stars mediocre, and 5-stars good. For these people, two motivations were at play. First, singling out 5-stars as the truly good category means that they wanted to maximize 5-stars. Second, putting 1- and 2-stars in one group suggests that they aimed to minimize the total counts of 1- and 2-stars as opposed to just minimizing1-stars on its own. Taken together, decision makers whose utility functions of the five-star rating scale followed a 1,2 — 3,4 —5 pattern would prefer options with more 5 stars, but not necessarily with fewer 1-stars. In our study, this was the feature of a more-variable product, and a tendency to choose a more-variable product made decision makers appear risk-seeking.

On the other end of this dimension, people assigned relatively higher values to Δ_{12} and Δ_{34}, and relatively lower values to Δ_{23} and Δ_{45} ($\Delta_{12} + \Delta_{34} > \Delta_{23} + \Delta_{45}$). Unlike their 1,2 — 3,4 — 5 counterparts, these people expanded the differences between 1- and 2- stars as well as 3- and 4-stars while compressed the differences between 2- and 3-stars as well as 4- and 5-stars, yielding a 1 — 2,3 — 4,5 pattern. They perceived 1-stars to be bad, 2- and 3-stars mediocre, and 4- and 5-stars good. For these people, two motivations were at play. First, perceiving 1-stars as the truly bad category means that they would want to minimize 1-stars. Second, grouping 4- and 5-stars together suggests that they would want to maximize their total counts as opposed to just maximizing 5-stars alone. Taken together, decision makers whose utility functions of the 5-star rating scale followed a 1 — 2,3 — 4,5 pattern would prefer options with fewer 1-stars, but not necessarily more 5-stars. In our study, this was the feature of a less-variable product, and a tendency to choose a less-variable product made decision makers appear risk-averse.

Dim 2. (Edge vs. Middle expanders) Explanation: This dimension contrasted stars on the two ends with those in the middle. Individuals on one end of this dimension ($\Delta_{12} + \Delta_{45} < \Delta_{23} + \Delta_{34}$) were middle expanders, distinguishing more between middle stars (2-4 stars). Meanwhile, they compressed the differences between edge stars; that is, they considered the utility of a 1-star to be similar to that of a 2-star, and the utility of a 4-star to be similar to that of a 5-star. Overall, this created the pattern of 1,2 — 3 — 4,5. This positioning means that middle expanders based their decisions not so much on the counts of 1-stars or 5-stars per se but on the total counts of 1- and 2-stars or 4-stars and 5-stars. In our experiment, when the average rating was extremely low (1.X), a more-variable product had a smaller total count of 1- and 2-stars and a greater total count of 4- and 5-stars than a less-variable product; when the average rating was high (4.X), however, a less-variable product had a smaller total count of 1- and 2-stars and a greater total count of 4- and 5-stars than a more-variable product. Hence, a goal of either minimizing the number of negative stars or maximizing the number of positive stars would lead middle expanders to be relatively risk-seeking when the average rating was extremely low by choosing a more-variable product and risk-averse when the average rating was high by choosing a less-variable product.

Conversely, decision makers on the other end of the spectrum were edge expanders who perceived 1-stars to be much worse than 2 stars and 5 stars to be much better than 4 stars ($\Delta_{12} + \Delta_{45} > \Delta_{23} + \Delta_{34}$). This created the pattern of 1 — 2,3,4 — 5. Since those who adopted this usage both disliked 1-stars and liked 5-stars to a great extent, they had to decide between minimizing 1-stars (which also means minimizing 5-stars) and maximizing 5-stars (which also means maximizing 1-stars). As such, their focus shifted as the average rating increased. When the average rating was extremely low (1.X), because there were considerably more negative stars than positive stars, edge expanders' decisions were primarily driven by their dislike of 1-stars. This led to risk-averse behavior, preferring a less-variable product with fewer 1-stars (and 5-stars) to a more-variable product with more 1-stars (and 5-stars). When the average rating was high (4.X), however, because there were considerably more positive stars than negative stars, edge expanders' decisions were primarily driven by their like of 5 stars. This 5-star-centered motivation reversed their risk choice behavior from risk-averse to risk-seeking, by which they preferred a more-variable product with more 5 stars (and 1 stars) to a less-variable product with less 5-stars (and 1-stars).

Dim 3. (Half amplifiers) Explanation: this dimension contrasted stars on the lower half of the five-star scale with those on the upper half. Individuals on one end of this dimension ($\Delta_{12} + \Delta_{23} < \Delta_{34} + \Delta_{45}$) assigned relatively lower values to Δ_{12} and Δ_{23} and higher values to Δ_{34} and Δ_{45}, perceiving the lower half of the five-star scale (1-, 2-, and 3-stars) to be similar utility-wise and expanded the differences between the upper half of the scale (4- and 5-stars). Decision makers on this end followed a 1,2,3 — 4 — 5 pattern, suggesting that while these people did not differentiate the lower ones that much, they especially cherished 5-stars. As such, this positioning made a more-variable product more desirable than a less-variable one, leading to risk-seeking behavior. This was because the former yielded a greater chance of getting a truly pleasant 5-star experience with its J-shaped distribution. On the other end of the dimension ($\Delta_{12} + \Delta_{23} > \Delta_{34} + \Delta_{45}$), individuals assigned relatively higher values to Δ_{34} and Δ_{45} and lower values to Δ_{12} and Δ_{23}, perceiving the upper half of the five-star scale (3-, 4-, and 5-stars) to be similar utility wise and expanded the differences between the lower half of the scale (1- and 2-stars). Decision makers on this end followed a 1—2 — 3,4,5 pattern, suggesting that while these people loathed 1-stars, they did not differentiate the higher ones that much. As such, this positioning made a less-variable product more desirable than a more-variable one, leading to risk-averse behavior. This was because the former yielded a lower chance of getting an awful 1-star experience with its relatively shorter 1-star-bar.

General Discussion

The rise of user-generated content has opened up many new opportunities for sellers and buyers to engage in economic exchanges. While online shopping has already become increasingly popular, the COVID-19 pandemic, in many ways, turned online shopping into a must activity. Since consumers had restricted access to many products during the pandemic, they had to rely increasingly on the experiences of others. While previous research has primarily focused on the influence of average rating and number of reviews on product choices, this research empirically explores the informational role of entire rating distributions. The impact of rating distributions on consumer behavior has only recently gained momentum, and no research has systematically examined whether and in what ways rating distributions influence consumer choice and, more importantly, what are the psychological causes of such preferences.

A combination of experimentation and cognitive modeling identifies three dimensions that separate how consumers interpret the five-star rating scale: 1) whether to compress or amplify the utility difference between middle (2-, 3-, and 4-stars) and extreme stars (1- and 5-stars), 2) whether 1-stars or 5-stars were singled out and whether 3-star ratings were deemed more positive or negative, and 3) which half of the rating scale (lower half: 1-, 2-, and 3-stars or upper half: 3-, 4-, and 5-stars) to differentiate more. While compressing the utility difference between middle and extreme stars (1,2---3---4,5) is similar to binary thinking reported by Fisher et al. (2018), there are other distinct, yet stable patterns across items in a single setting among decision makers, including edge expanders and lower-half expanders. While it is not surprising that people ascribe different meanings to the five-star rating scale typically implemented by reputation and feedback systems, this work contributes to the current understanding of scale usage and online consumer decision-making by experimentally showing the different patterns people exhibit. Different variations in these patterns can lead to both risk-seeking and risk-averse behavior, at different points in the value spectrum. This work used phone cases as the target, and while it is not certain whether these observed patterns transcend products, this serves as a possible starting point for future analysis of these trends. Because of these systematic differences, this work may also help to reconcile the mixed findings on the influence of rating variance.

An important extension of this research to the previous literature on the informational role of rating variance is that this research takes individual differences into consideration. In the current study, we observed groups of people, such as middle expanders, who preferred high variance when the valence was on the lower end and low variance when the valence was on the higher end. This, however, was not the full story. In addition to people who behaved similarly to what previous studies (e.g., Sun, 2012) have suggested, we also observed people who behaved exactly opposite: They appeared to prefer less-variable products when the valence level was on the lower end and more-variable products when the valence was on the higher end. There were also groups of people whose patterns of risky choice behavior were more consistent across different levels of valence. These individual differences were strong enough that simply focusing on the aggregate pattern would be misleading.

It should be noted that this paper has illustrated the impact of rating variance in people's decision making by considering only the cases when there are no differences between average ratings and number of reviews. Future research can look into how people trade off between rating valence and rating variance. It might be intuitive to say that people would prefer the option with higher rating valence. However, when decision makers have strong needs for uniqueness or self-expression, polarization in opinions has been consistently preferred over uniformity (Rozenkrants et al., 2017). High dispersion in ratings also makes a product more desirable when high heterogeneity in tastes is expected (e.g., painting) (He & Bond, 2015). Thus, the answer to whether people will still pay attention to rating variance when differences in average ratings are obvious may be more complicated than one would have expected.

In an era where personalization is highly emphasized, realizing how people's different interpretations of the rating scale can lead to drastically different preferences for products is key to enhancing consumption experiences as well as product sales.

References

Amendola, L., Marra, V., & Quartin, M. (2015). The evolving perception of controversial movies. *Palgrave Communications, 1(1),* 1-9.

Arndt, J (1967). *Word of Mouth Advertising.* New York: Advertising Research Foundation, 1967.

Caro, N. (Director). (2020). *Mulan* [Film]. Walt Disney Pictures.

Chevalier, J. A., & D. Mayzlin. (2006). The effect of word of mouth on sales: Online book reviews. *J. Marketing Res. 43(3)* 345–354.

Fischhoff, B., & Broomell, S. B. (2020). Judgment and decision making. *Annual review of psychology, 71, 331-*355.

Floyd, K., Freling, R., Alhoqail, S., Cho, H. Y., & Freling, T. (2014). How online product reviews affect retail sales: A meta-analysis. *Journal of Retailing, 90(2),* 217-232.

He, S. X., & Bond, S. D. (2015). Why is the crowd divided? Attribution for dispersion in online word of mouth. *Journal of Consumer Research, 41(6),* 1509-1527.

Hu, N., Pavlou, P. A., & Zhang, J. (2006). Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of online word-of-mouth communication. In *Proceedings of the 7th ACM conference on Electronic commerce* (pp. 324-330).

Hu, N., Zhang, J., & Pavlou, P. A. (2009). Overcoming the J-shaped distribution of product reviews. *Communications of the ACM, 52(10),* 144-147.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–283.

Khare, A., Labrecque, L. I., & Asare, A. K. (2011). The assimilative and contrastive effects of word-of-mouth volume: An experimental examination of online consumer ratings. *Journal of Retailing, 87(1),* 111-126.

Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science, 62(12),* 3412-3427.

Powell, D., Yu, J., DeWolf, M., & Holyoak, K. J. (2017). The love of large numbers: A popularity bias in consumer choice. *Psychological science, 28(10),* 1432-1442.

Rosario, A. B., Sotgiu, F., De Valck, K., & Bijmolt, T. H. (2016). The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *Journal of Marketing Research, 53(3),* 297-318.

Rozenkrants, B., Wheeler, S. C., & Shiv, B. (2017). Self-expression cues in product rating distributions: When people prefer polarizing products. *Journal of Consumer Research, 44(4),* 759-777.

Sun, M. (2012). How does the variance of product ratings matter? *Management Science, 58(4),* 696-707.

West, P. M., & Broniarczyk, S. M. (1998). Integrating multiple opinions: The role of aspiration level on consumer response to critic consensus. *Journal of Consumer Research, 25(1),* 38-51.