Online Decision Making with Icon Arrays

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

Abstract

Leveraging people's proficiency in extracting summary statistics from ensembles, we conducted two studies in which we presented rating information from consumer feedback systems through color-coded icon arrays. The investigation aims to explore how different icon array arrangements (ascending, descending, random) influence decision making and average estimation across varying levels of ensemble means (average ratings) and ensemble sizes (review volume). Our results revealed four key insights: 1) Preferences for rating variance differed, particularly at the extremes of the average rating spectrum when ensembled were dominated by one or two rating categories. 2) Structured information yielded greater certainty in responses, with confidence increasing when the task setup aligned with task goals. 3) Ensemble size prompted individuals to adapt strategies based on contextual needs. 4) Unstructured presentations led to higher estimation accuracy, suggesting that a lack of structure may encourage heightened processing effort.

Keywords: Icon arrays, ensemble perception, consumer feedback, decision making, estimation

Introduction

Information visualization significantly influences decision outcomes. Depending on how data is visually represented, the same dataset can yield various interpretations. Different visualizations highlight distinct aspects of the data, offering comparative advantages for tasks such as option comparison and trend forecasting (Hawley et al., 2008; Tait et al., 2010; Yuan et al., 2019).

Among numerous applied contexts, online shopping platforms with consumer feedback (e.g., ratings and reviews) provide familiar environments for decision making based on data. Despite advancements in consumer feedback systems, the visual representation of quantitative data has remained largely unchanged. Current systems typically use bar graphs to depict rating profiles, with variations in how counts of each star rating category are conveyed. Some platforms present raw numbers (e.g., Tripadvisor), some use percentages (e.g., Amazon), and others show no numbers (e.g., Google, Yelp).

Considering icon arrays' demonstrated ability to aid individuals' comprehension of probabilistic information by leveraging the human perceptual system (Nelson et al., 2008), particularly those with lower numeracy skills (Galesic et al., 2009), we are intrigued by the potential impact of incorporating icon arrays in the context of online shopping with consumer feedback. An additional advantage of icon arrays is their perceived trustworthiness compared to other visual and non-visual forms (Hawley et al., 2008). This aligns well with the consumer feedback space, where concerns about the credibility and authenticity of information are growing, particularly due to the proliferation of fake reviews. Moreover, icon arrays, composed of individual elements, can be used to communicate ordinal and temporal information, offering a unique perspective that goes beyond traditional bar graphs used in consumer feedback systems. This is akin to how sites offer different sorting methods.

The proposed approach seems promising, but a crucial question is whether individuals possess the capacity to process rating information presented as icon arrays and differentiate between distinct star rating categories coded by colors (similar to Yelp's display). Insights from ensemble perception literature suggest that viewers indeed have the capacity. Humans demonstrate an extraordinary ability to extract various summary statistics from a group of elements (Alvarez, 2011), spanning low-level stimuli like hue and brightness, mid-lese stimuli like size, to high-level stimuli such as facial expressions. Beyond averages, people are also able to extract variance information from different ensembles (see Whitney & Leib, 2018 for a review).

For color ensemble perception specifically, individuals can effectively extract summary statistics for means and variance (Maule et al., 2014). Hence, adopting an ensemble format with color-coded tokens to represent distinct star rating categories offers the potential to convey four pieces of information: 1) average ratings, 2) rating variance, 3) the number of reviews, and 4) sorting information. Leveraging an ensemble format for the number of reviews allows for direct manipulation of visual impact by displaying more or fewer icons. For instance, a sample size of 100 and a sample size of 999 may share the same digit count (3), potentially diminishing the salience of size impact. However, ensembles, with their distinct visual areas, highlight the contrast between 100 and 999 tokens (assuming identical token sizes). Regarding sorting information, the common methods enabled by rating sites include "highest," "lowest," and "most recent." This information could be communicated by organizing color-coded tokens in specific arrangements, reminiscent of Xiong et al.'s (2022) investigation into the impact of different arrangements of icon arrays on percentage estimation.

The present study investigates how perceptual features such as icon arrangement and color variation influence preferences and judgments in online shopping. Specifically, we employed color ensembles to convey rating information, aiming to explore how icon arrangements (ascending from lowest to highest star ratings, descending from highest to lowest star ratings, and random) influence decision making and average estimation across different ensemble means (representing average ratings) and ensemble sizes (representing the number of reviews).

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Experiment 1

Experiment 1 explores how different arrangements of color-coded icon arrays influence people's preferences for rating variance and their confidence in those preferences.

Method

Participants. In exchange for course credit, 108 undergraduate students from a major university in the Midwest participated in the experiment.

Materials and Design. The experiment was a within-subject design manipulating three Vs of online reviews: valence (average ratings), volume (number of reviews), and variance (dispersion in opinions). For valence, we introduced four levels to cover the entire spectrum of a 5-star rating scale: extremely low (1.X), low (2.X), medium (3.X), and high (4.X). For volume, we included three levels: low (< 50 number of reviews), medium (50 – 400 number of reviews), and high (> 400 number of reviews). For simplicity and differentiation purposes, the rest of the paper used the valence level of 1.X, 2.X, 3.X, and 4.X and the volume level of low, medium, and high. At each combination of valence and volume level, we chose three variations. For example, at the combination of a valence level of 1.X and a low volume level, we introduced 1.5 (50), 1.7 (45), and 1.9 (30). We created three products for each combination of valence and volume level, leading to three pairwise comparisons. This led to a total of 108 trials: 4 (valence level) x 3 (volume level) x 3 (variations) x 3 (pairwise comparisons). To-be-compared products were presented to participants in pairs with identical valence and volume levels. The only difference between the two products in each pair was their rating variance.

We used the low volume level as the basis of stimuli design. For medium and high-volume levels, we simply retained star proportions by applying a multiplier across the board. For example, at one of the low volume levels, we had 1.5-star ratings out of 50 stars with 42 1-star ratings, one 2-star rating, one 3-star rating, two 4-star ratings, and four 5-star ratings. When expanding this into the level of high volume with 450 stars, we simply multiplied every star rating category with 9, resulting in 378 1-star ratings, nine 2-star ratings, nine 3-star ratings, 18 4-star ratings, and 36 5-star ratings.

An ensemble format was used to display the statistical characteristics of product ratings. The colors of different star ratings were modeled after those displayed on Yelp.com. The purpose of introducing three variations at each level of valence and volume combination was that we could display star ratings in one of the following three arrangements: ascending (from 1-star ratings to 5-star ratings), descending (from 5-star ratings to 1-star ratings), and random. This manipulation was referred to as display order in the rest of the paper. On any given trial, the display was the same for each product pair. The overall experiment featured eight trials of ascending, descending, and random orders, respectively. Figure 1 shows a sample of each display order.

Procedure. Trials were randomized and counterbalanced. Participants were tasked to indicate their purchase preferences on a 6-point Likert scale, ranging from "*1* = definitely buy the left product" to "*6* = definitely buy the right product."

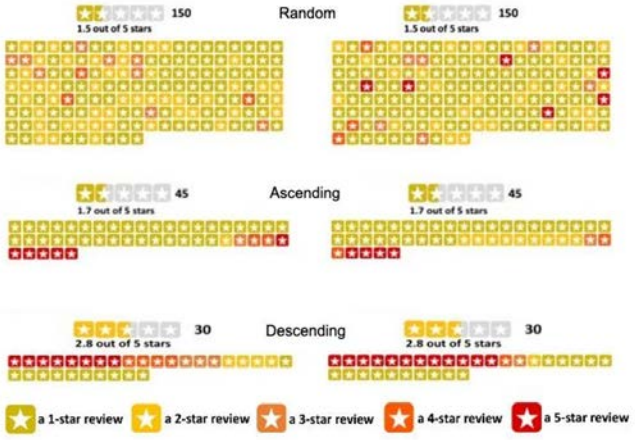

Figure 1. Sample trials of three display orders.

Results

Preference in Rating Variance To examine participants' preferences for rating variance when presented in an ensemble format, we recoded their responses as binary choices. Values of 1 represented responses within the range of 1-3 when the left product had a lower rating variance, or responses within the range of 4-6 when the right product had a lower rating. Values of 0 represented responses within the range of 1-3 when the left product had a higher rating variance, or responses within the range of 4-6 when the right product had a higher rating variance. As responses were recoded into 0 or 1, we created the probability of picking a less-variable product as a new dependent variable.

A three-way within-subjects ANOVA analysis was conducted with display order, valence level, and volume level as independent variables and the probability of picking a less-variable product as a dependent variable. While there was a significant three-way interaction between the three factors, $F(12,1284) = 1.88$, $p = .033$, post-hoc analyses showed that there were no differences between the effects of ascending and descending orders, $t(107) = -1.27$, $p = .207$. Therefore, we decided to combine ascending and descending orders into a new display order: structured. The rest of the results section on Task 1 would report comparisons between structured and unstructured display orders on participants' probability of picking a less-variable product.

We conducted a new three-way within-subjects ANOVA analysis to reflect the combination of ascending and descending orders into an overall structured order. As shown in Figure 2, there was a significant three-way interaction between display order, valence level, and volume level, $F(6,642) = 2.811$, $p = .001$. Simple two-way interactions at each level of valence level showed that there was a significant

interaction between volume level and display order at the valence level of 4.X, $F(1.86,200) = 4.79$, $p = .011$. Post-hoc analyses with Bonferroni adjustments showed that when the valence level was 4.X, the probability of picking a less-variable product between a structured and unstructured order differed when the volume level was low, $F(1,107) = 17.7$, $p < .001$, or medium, $F(1,107) = 5.6$, $p = .02$. The probability remained the same when the volume level was high, $F(1,107) = 0.347$, $p = .557$. There was a simple main effect at the valence level of 1.X, $F(1,107) = 19.6$, $p < .001$, 2.X, $F(1,1.07) = 4.53$, $p = .036$, and 4.X, $F(1,107) = 13.5$, $p < .001$. Post-hoc analyses revealed that the effect of display order was significant in six combinations of valence and volume level, reported in the valence level – volume level format: 1.X – low ($p < .001$), 1.X – medium ($p = .042$), 1.X – high ($p = .01$), 2.X – low ($p = .008$), 4.X – low ($p < .001$), 4.X – medium ($p = .02$). Across all these cases, structured orders led to a significantly higher probability of picking a less-variable product than unstructured orders.

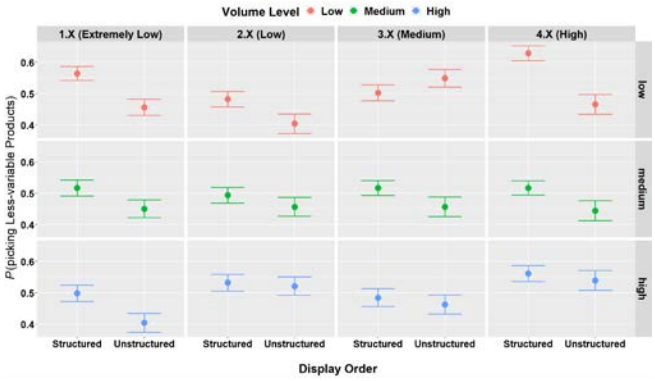

Figure 2. The effect of display order (structured vs. unstructured) on preference for rating variance by valence level and volume level.

Simple two-way interactions were observed at each volume level, revealing a significant simple two-way interaction between valence level and display order specifically at low volume levels, $F(3,321) = 7.58$, $p < .001$. Post-hoc analyses with Bonferroni adjustments revealed that at the low volume level, the probability of choosing a less-variable product varied between two display orders across three valence levels: 1.X, 2.X, and 4.X. For all these valence levels, structured orders were associated with significantly higher probabilities of picking less-variable products compared to unstructured orders, ps $< .001$. In the three significant valence-level combinations, the probabilities of picking lower rating variance between structured and unstructured orders were as follows: 1.X – low – structured: $M = 0.56$, $SD = 0.23$ vs. 1.X – low-unstructured: $M = 0.46$, $SD = 0.27$; 2.X – low – structured: $M = 0.48$, $SD = 0.26$ vs. 2.Xm – low – unstructured: $M = 0.40$, $SD = 0.32$; 4.X – low – structured: $M = 0.63$, $SD = 0.25$ vs. 4.X – low – unstructured: $M = 0.47$, $SD = 0.33$.

Simple two-way interactions at each type of display order indicated a significant two-way interaction between valence level and volume level with both structured, $F(6,642) = 3.82$, $p < .001$, and unstructured orders, $F(6,642) = 4.27$, $p < .001$. Through pairwise comparisons with Bonferroni adjustments, we identified five significant differences (ps: $< .001$— $.047$) in volume levels across various combinations of valence level and display order. Two major findings emerged from these patterns. First, more differences in preference toward rating variance were observed across volume levels when tokens were presented in an unstructured fashion. Second, the primary disparities originated from comparisons involving the low or high volume level.

We further examined probabilities of picking less-variable products against a threshold of 0.5, seeking to understand whether participants displayed a genuine preference for a particular type of rating variance, or if their choices were of a 50-50 split (i.e., no preference). To this end, multiple one-sample t-tests against the 0.5 threshold were conducted. Five conditions where a significant preference emerged, presented in the format of display – valence level – volume level. Among these, three exhibited a preference for lower rating variance, with probabilities of picking less-variable products above 0.5: structured – 1.X – low ($p < .001$), structured – 4.X – low ($p < .001$), structured – 4.X – high ($p = .018$). In contrast, two conditions indicated a preference for higher rating variance, with probabilities of picking more-variable products below 0.5: unstructured – 1.X – high (p = .0019), unstructured – 2.X – low ($p = .0026$). Additionally, two conditions were marginally below 0.5, suggesting a tendency towards more-variable products: unstructured – 1.X – low ($p = .097$) and unstructured – 1.X – medium ($p = .086$). Taken altogether, at the valence level of 1.X, there was a distinct preference for more-variable products under unstructured presentations, evident in the consistently lower-than-0.5 average probabilities of choosing less-variable products: $M = 0.46$, $SD = 0.27$ for low volume, $M = 0.45$, $SD = 0.30$ for medium volume, and $M = 0.40$, $SD = 0.31$ for high volume.

Choice Confidence We conducted a similar three-way within-subjects ANOVA analysis with participants' response extremity as a dependent variable, which was used as a proxy of choice confidence. To determine this metric, we calculated the absolute differences between the Likert scale responses and 3.5 (the midpoint of the scale).

As illustrated in Figure 3, there was a significant three-way interaction between valence level, volume level, and display order. Given our research questions, we were particularly interested in terms pertinent to display order. Simple two-way interactions revealed that there was a significant interaction between volume level and display order at the valence level of 2.X, $F(1.89, 202) = 11.3$, $p < .001$, 3.X, $F(2, 214) = 5.34$, $p = .005$ and 4.X, $F(2,214) = 6.48$, $p = .002$. Post-hoc analyses with Bonferroni adjustments showed that structured orders consistently led to higher choice confidence than unstructured orders across all volume levels when average ratings were moderately good (3.X and 4.X), ps $< .001$. This trend extended to valence-volume combinations featuring a valence level of 2.X and a medium or large

volume size, *p*s < .001. However, for every volume size at the valence level of 1.X and a low volume size at the valence level of 2.X, there was no significant difference in choice confidence between structured and unstructured orders, *p*s > .1. Choice confidence was significantly higher at the valence of 4.X than the other three valence level, *p*s < .001.

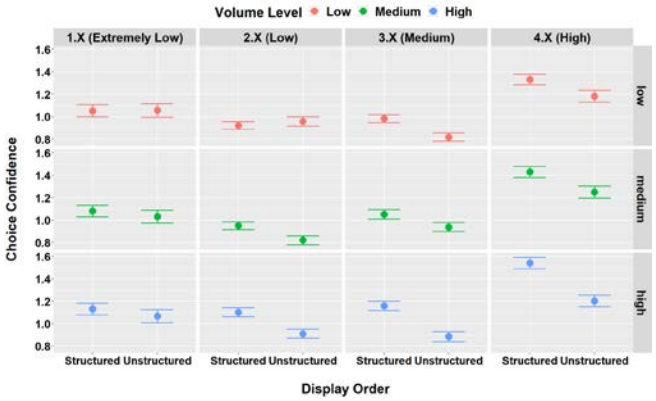

Figure 3. The effect of display order on choice confidence by valence and volume level.

Overall, there was a main effect of display order, $F(1, 107) = 50.91$, $p < .001$, in which choice confidence was significantly higher when tokens were presented with structured ($M = 1.14$, $SD = 0.32$) than unstructured orders ($M = 1.01$, $SD = 0.34$).

Discussion

One of the most consistent findings is that at the valence level of 1.X, unstructured presentations resulted in a preference for larger rating variance. We speculate that this was due to the distinct star composition. At a valence level of 1.X, there were a lot more negative star ratings than positive star ratings, with a substantial presence of 1- and 2-star ratings and some limited presence of 4- and 5-star ratings.

Structured presentations facilitated various visual cues, allowing participants to trace the number of 5-star ratings or 1-star ratings. They could even engage in intra-product (calculate net star counts) or inter-product comparisons (determine which side had a higher count of particular star counts). The environmental affordance of structured orders might have played a role in shaping participants' strategies.

In contrast, unstructured orders posed challenges for direct comparisons at the valence level of 1.X. Side-by-side comparisons of poorly rated products conflicted with the selection goal. Individuals needed to rationalize their choices with positive aspects (albeit limited). As a result, counting seemed the most straightforward approach, especially with fewer positive star ratings at the valence level. In our experiment, a more-variable product had a higher number of positive star ratings (whether considering 5-star only or a combination of 4-star and 5-star ratings). As such, regardless of whether participants perceived only 5-star ratings as positive or lumped 4-star and 5-star ratings together as positive (Yu et al., 2022; Fisher et al., 2018), their desire to seek products with more positive signals led to a preference for more-variable products. While participants could use a similar approach with structured orders, the salient structure provided flexibility, possibly encouraging quick estimation over precise counting.

On the opposite side of the spectrum, where average ratings were above 4 stars, we observed a preference for less-variable products with structured orders across volume levels. We theorize that this was because at the valence of 4.X, products were attractive, despite the exact composition of rating profiles. When choosing between two desirable products, negative information becomes more diagnostic, leading to a focus on the number of negative stars. As the less-variable products were those with fewer 1-star ratings or a fewer total of 1-star and 2-star ratings, whether participants considered only 1-star ratings as negative or perceived both 1-star and 2-star ratings as negative (Yu et al., 2022; Fisher et al., 2018) did not make a difference. Structured presentations facilitated such comparisons.

Why do individuals prioritize positive ratings at the valence level of 1.X with structured orders, but not negative ratings at the valence level of 4.X? Intuitively, one might expect people to focus on the diagnostic category with the fewest stars in both cases. We speculate that this distinction stems from people's motivations. With an average rating above 4 stars, both products were attractive, creating a "one cannot go wrong with either opinion" situation. Consequently, there might be less motivation to pick one from two already desirable options. Since there are no right or wrong answers with preferences, this may be especially the case when rating information does not have a salient structure. This lack of effort aligns with the observation that choices were essentially evenly split with unstructured orders at the valence of 4.X.

There is also the possibility that specific strategies and response patterns were influenced by perceptual capabilities. When the valence levels were 2.X or 3.X, the positive and negative star ratings were evenly distributed, making it challenging to discern differences when represented with colored tokens. This difficulty in visual discrimination could explain the limited observed differences at these valence levels. In contrast, when the valence levels were 1.X or 4.X, the ensembles were predominantly occupied by one or two rating categories, creating significant contrasts in the size of color patches. Given our experimental design, at the valence level of 1.X, the presentation mostly consisted of green and yellow star arrays with subtle hints of orange and red. At the valence level of 4.X, the presentation was mostly composed of orange and red stars with glimpses of green and yellow.

Surprisingly, or not, the impact of volume level was limited. Upon deeper examination, the main differences in probabilities of picking less-variable products arose from comparisons that included either the low- or high-volume level. While one could speculate that participants acted as intuitive statisticians, adjusting their decisions based on presentation size, we attributed these observations to a shift in strategies stemming from the stimuli interface.

Specifically, when the volume level was low (< 50 tokens), participants could easily count stars and make decisions with precise numbers. However, with a large volume level (300 – 400 tokens), participants may have been more inclined to resort to ensemble perception to form an overall impression of a rating profile. As single-item recognition is not a prerequisite of ensemble coding (Whitney & Leib, 2018), it became especially suitable at the high-volume level in our study. This implies that participants did not need exact counts or even to identify the meaning of each token to gauge the homogeneity or heterogeneity in previous buyers' opinions. Moreover, the summary representation of rating distributions independently influences people's subjective evaluations even when average ratings are presentations (Fisher et al., 2018). As a result, participants could have arrived at different conclusions about the attractiveness of a product, even when average ratings were provided during our experiment. In short, to obtain information on rating variance, individuals can either count, estimate, or employ alternative approaches. The appeal of various strategies changes at different stages, influenced by differences in volume levels.

Examining participants' response extremity revealed a consistent trend of greater confidence in decision making with structured presentations compared to unstructured presentations. This suggests that structured information affords greater certainty. A similar idea was expressed by Brannon and Carson (2003), who observed that both nurse experts and novices reported greater certainty in their diagnosis with information that is high in structure compared with information low in structure.

The absence of differences in choice confidence at the valence level 1.X suggests an intriguing nuance. Rather than being influenced by icon arrangements, this phenomenon might be due to our experimental design. In everyday scenarios, individuals typically consider purchasing products with a minimum average rating of 3.3 stars (BrightLocal, 2023). Thus, choosing between two equally unappealing options with average ratings as low as 1.X stars may have influenced participants' moderately held attitudes overall.

Another consistent finding is that individuals were more confident when the valence level was 4.X compared to other levels. Choosing between two highly rated products aligns with a selection mindset, in line with our task goal. Products rated four stars and above are perceived as more appealing, boosting confidence in purchasing intentions, regardless of the final choice.

In summary, structured information affords greater certainty in answers, and this confidence increases when the task setup aligns with task goals. The influence of ensemble size becomes evident as it prompts individuals to adapt strategies that best suit their contextual needs.

Experiment 2

Experiment 1 asked participants to indicate their preferences, so there were no right or wrong answers. While we demonstrated that people were more confident in their choices with structured presentations, we did not directly assess the accuracy associated with different display orders. Hence, Experiment 2 required participants to evaluate two representations and determine which side had a higher average rating, introducing right or wrong answers.

Method

Participants. The same 108 participants from Experiment 1 participated in Experiment 2.

Materials and Design. The number of icons in Experiment 2 was restricted to 100 to ensure the manageability of the task, aiming to simplify participants' estimation of percentages. The experimental design manipulated two variables: valence level and display order. For the valence level, similar to Experiment 1, valence featured four levels: extremely low (1.X), low (2.X), medium (3.X), and high (4.X). For display order, three display orders were featured – ascending, descending, and random – each with eight trials. Within each display order, two comparisons were presented for each valence level: *Same comparisons* involved products with both higher or lower average ratings and variances. For instance, a comparison might feature a 4.2-star product with 20 1-star ratings and 80 5-star ratings (SD = 1.61) against a 4-star product with 100 4-star ratings (SD = 0). *Different comparisons* involved products with a higher average rating paired with a lower variance or vice versa. For instance, a 4.5-star product with 50 4-star ratings and 50 4-star ratings (SD = 0.50) versus a 4-star product with 25 1-star ratings and 75 5-star ratings (SD = 1.74). This manipulation was referred to as trial type in the rest of the paper.

Procedure. Participants faced 24 trials where they were required to decide which side of a product pair had a higher average rating through two-alternative forced-choice tasks.

Results

We conducted a three-way within-subjects ANOVA with valence level, display order, and trial type as independent variables and participants' response accuracy as a dependent variable. As there were no differences in accuracy between an ascending and descending order, $p > .1$, we decided to combine descending and ascending orders into the category of structured orders. With this combination, we conducted a new three-way within-subjects ANOVA to reflect this change. The three-way interaction was not significant, $F(3,321) = 0.044$, $p = .99$.

Figure 4 showcases a significant two-way interaction between valence level and display order, $F(3,321) = 3.777$, $p = .011$. Simple main effects were conducted at each level of the valence level to further explore the effect of display order on response accuracy. Results showed that the effect of display order was significant at the valence level of 4.X, $F(1,107) = 7.89$, $p = .006$. At this valence level, structured orders (M = 0.56, SD = 0.24) led to significantly lower accuracy than unstructured orders (M = 0.67, SD = 0.33).

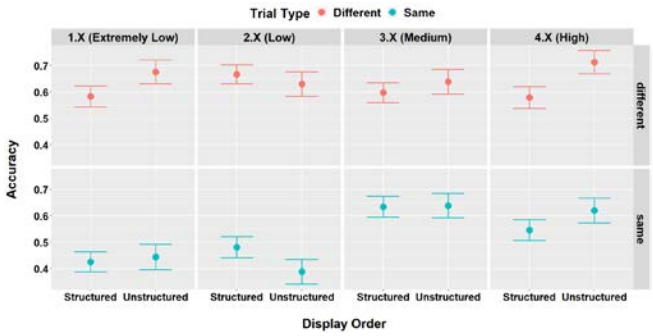

Figure 4. Accuracy by display order, trial type, and valence level. Overall, estimation accuracy was higher with unstructured than structured presentations.

A significant two-way interaction between valence level and trial type was observed, $F(3,321) = 6.586$, $p < .001$. Simple main effects were conducted at each level of the valence level to further explore the effect of trial type on response accuracy. Results showed that the effect of trial type was significant at the level of 1.X and 2.X (lower end of the valence spectrum), ps $< .001$. In both cases, different comparisons resulted in significantly higher estimation accuracy compared to same comparisons. Specifically, at the valence level of 1.X, the average accuracy was 0.61 (SD = 0.36) for different comparisons, and 0.43 (SD = 0.34) for same comparisons. Similarly, at the valence level of 2.X, the average accuracy was 0.65 (SD = 0.34) for different comparisons and 0.45 (SD = 0.37) for the same comparisons. In summary, display order influenced accuracies differentially at the lower end of the valence spectrum.

We also observed a main effect of valence level. $F(3,321) = 7.633$, $p < .001$. Post-hoc analyses, adjusted with Bonferroni corrections, revealed three significant comparisons. The accuracy at the valence level of 1.X (M = 0.52, SD = 0.17) was significantly lower than both at the valence level of 3.X (M = 0.63, SD = 0.20), $p < .001$ and 4.X (M = 0.59, SD = 0.20), $p = .006$. In addition, the accuracy at the valence of 2.X (M = 0.55, SD = 0.17) was significantly lower than at the valence level of 3.X (M = 0.62, SD = 0.20), $p = .003$. Overall, there is some suggestion that accuracy improves with increasing ensemble means.

Discussion

The results of Experiment 2 indicate that display orders directly influenced individuals' estimation accuracy. The direction of the finding is especially intriguing – unstructured ensembles lead to higher accuracy in average rating judgments than structured ensembles. This contradicts the findings in Xiong et al.'s (2022) work, where top (bottom), row, and diagonal arrangements resulted in overall higher accuracy in people's percentage estimates compared to arrangements such as edge, central, and random. Aside from the differences in the number of distinct token categories used (Xiong et al., had black and white while we used five), we posit that the higher accuracy associated with unstructured arrangements in our experiment is due to increased cognitive effort. Ascending or descending orders may give people the impression of easiness, similar to the perceptual fluency effect (Jacoby & Dallas, 1981). With unstructured presentations, the task may have seemed challenging to individuals, prompting them to actively engage in counting, calculation, or other cognitive processes to arrive at conclusions. Our findings indicated that creating an environment requiring a reasonable level of effort can foster engagement and result in higher performance. The advantage of an unstructured order was particularly pronounced when options were attractive, as indicated by the valence level of 4.X. As mentioned previously, this may be attributed to the alignment between the experimental setup (attractive product pairs) and the task goal (selection). Unlike Experiment 1, where preferences were solicited and people could not go wrong with either option at the valence level of 4.X, Experiment 2 emphasized accuracy. The absence of structure could further amplify motivation to engage in deliberate processes to get correct answers. Future research could explore whether people tend to overestimate or underestimate average ratings across different types of display orders and levels of valence.

Another interesting finding is that different comparisons led to significantly higher accuracy compared to same comparisons. In different comparisons, a higher (lower) average rating is paired with a lower (higher) rating variance, whereas in same comparisons, a higher (lower) average rating is paired with a higher (lower) rating variance. The challenge posed by high-variance color ensembles makes the extraction of summary statistics more difficult and less accurate (de Gardelle & Summerfield, 2011). This suggests that rating variance can distort the perception of the actual average rating. Our exploratory item analysis indicated a discernable trend: higher variance tends to make a higher average rating appear lower than it actually is, while making a low average rating seem higher than it truly is. Future research could explore the directions in which variance influences the estimation of means across varying levels of means with structured vs. unstructured ensemble formats.

Conclusion

The present study examined the effectiveness of icon arrays in conveying consumer ratings in online decision-making contexts. It presents two experiments investigating how three different color-coded icon arrays – ascending, descending, and random – influence users' decision-making, confidence, and estimation accuracy regarding product ratings. We found that 1) preferences differed for rating variance, 2) information structure affected choice confidence, 3) the ensemble size played a role in strategy adaptation, and 4) unstructured presentations led to higher estimation accuracy (perhaps because they evoked more careful processing).

We encourage future studies to further investigate whether these findings are specific to the consumer feedback context, given its relatively recent establishment.

References

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in cognitive sciences*, *15*(3), 122-131.

Brannon, L. A., & Carson, K. L. (2003). The representativeness heuristic: influence on nurses' decision making. *Applied Nursing Research*, *16*(3), 201-204.

BrightLocal (2023), Local Consumer Review 2023. https://www.brightlocal.com/research/local-consumer-review-survey/

De Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences*, *108*(32), 13341-13346.

Fisher, M., Newman, G. E., & Dhar, R. (2018). Seeing stars: How the binary bias distorts the interpretation of customer ratings. *Journal of Consumer Research*, *45*(3), 471-489.

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: overcoming low numeracy. Health psychology, 28(2), 210.

Hawley, S. T., Zikmund-Fisher, B., Ubel, P., Jancovic, A., Lucas, T., & Fagerlin, A. (2008). The impact of the format of graphical presentation on health-related knowledge and treatment choices. *Patient education and counseling*, *73*(3), 448-455.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, *110*(3), 306.

Maule, J., Witzel, C., & Franklin, A. (2014). Getting the gist of multiple hues: Metric and categorical effects on ensemble perception of hue. *JOSA A*, *31*(4), A93-A102.

Nelson, W., Reyna, V. F., Fagerlin, A., Lipkus, I., & Peters, E. (2008). Clinical implications of numeracy: theory and practice. Annals of behavioral medicine, 35(3), 261-274.

Tait, A. R., Voepel-Lewis, T., Zikmund-Fisher, B. J., & Fagerlin, A. (2010). Presenting research risks and benefits to parents: does format matter? *Anesthesia and analgesia*, *111*(3), 718.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual review of psychology*, *69*, 105-129.

Xiong, C., Sarvghad, A., Goldstein, D. G., Hofman, J. M., & Demiralp, Ç. (2022, April). Investigating perceptual biases in icon arrays. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).

Yu, J., Landy, D., & Goldstone, R. (2022). How Do People Use Star Rating Distributions? In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, No. 44).

Yuan, L., Haroz, S., & Franconeri, S. (2019). Perceptual proxies for extracting averages in data visualizations. *Psychonomic bulletin & review*, *26*, 669-676.