User-Generated Star Ratings Are Not Inherently Comparative

https://cuboulder.zoom.us/j/9541401133

Matt Meister & Nicholas Reinholtz

Summary

We demonstrate an intrinsic problem with the use of star ratings in comparative choice.

Star ratings are produced in isolation (non-comparatively) but are used comparatively. Because between-subjects (isolated) judgments can lead to illogical rating patterns (Birnbaum 1999) when raters do not share frames of reference, relative differences in ratings may not reflect relative differences in quality. We demonstrate how **two extremely similar products that differ in expectations may result in illogical ratings patterns.**

This simple empirical demonstration illustrates how the **structural misalignment** between ratings' procurement and use can lead consumers to choose suboptimal products.

Method

We offered two tasks to AMT workers. Both tasks required participants to complete the same – 10 trials of counting the number of 0s in a 6x6 grid (below):

Complete the Task

Count the number of zeros

0	1	0	0	1	0
0	1	0	1	0	1
0	1	0	0	1	0
1	1	0	1	1	0
0	1	1	0	0	0
0	1	1	0	0	1

Tasks only differed in bonus payments. Both had a 90% chance of paying 5¢ per correct answer. The "better" task had a 10% chance of paying 25¢. The "worse" had a 10% chance of paying 4¢.

	Better task	Worse task
Likely Bonus Rate (90%)	5¢	5¢
Unlikely Bonus Rate (10%)	25¢	4¢
Expected Bonus Rate	7¢	4.9¢

University of Colorado **Boulder**

Method (cont'd)

In Study 1 – Phase 1, workers were randomly placed into one of the two tasks. They learned about the task, its possible payments (and their probabilities), then counted 0s and were paid. They finished by rating the task on a 1-5 star scale. Those in the objectively better task provided lower average ratings (3.73 vs 4.46) because they were paid at the bottom of their frame of reference.

We used these ratings as stimuli for future participants. In Study 1 – Phase 2, participants either saw Star Ratings, Pay Information, or Both. Below is the information presented for the Both condition.

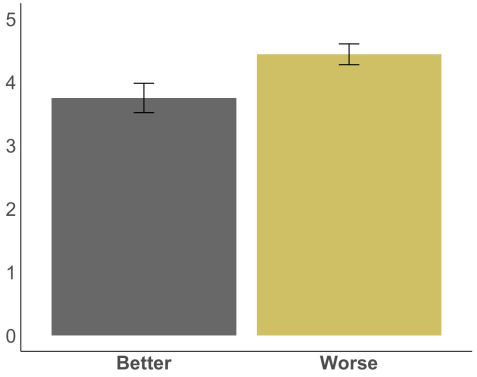
	Task A	Task B
Number of Trials	10	10
Minimum pay per correct response	4¢	5¢
Maximum pay per correct response	5¢	25¢
Average payment	45¢	66¢
Average rating	★★★★½ 4.46/5	★★★★☆ 3.73/5

In task A, 90% of people received 5¢ per correct, and 10% 4¢.
In task B, 90% of people received 5¢ per correct, and 10% 25¢.

Hypotheses & Results

1. In isolated evaluation, an inferior alternative may receive higher ratings than its superior when it engenders higher expectations

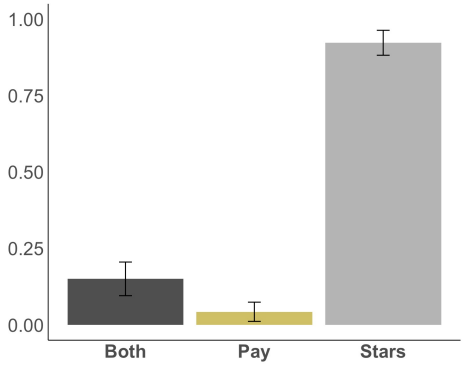
- Support in S1-P1: $F(1, 199) = 22.91$, $d = .68$



Hypotheses & Results

2. The **mere presence of star ratings** will lead workers to be more likely to select the utility-minimizing task.

- Support in S1-P2



Secondary Hypotheses

3. The presence of text reviews will not "fix" H2

- Text reviews in S3 **increased** poor choices: ($M_{TextReviews}$ = 73%, $M_{NoReviews}$ = 69%, z = -1.84, p = .066)

4. Participants do not actually enjoy the worse task more

- 82% of S2 participants chose to repeat the objectively better task ($\chi^2 = 21.41$, $p < .001$) after first completing both

Conclusion

Three studies demonstrate an intrinsic problem with the use of star ratings in comparative choice.

- Because star ratings are produced in isolation, they need not be comparable across alternatives.
- When products reliably engender different expectations, we can expect their ratings to be less comparable (Oliver 1980)
- Consumers do not anticipate this – they think that ratings are comparable. As a result, they are susceptible to make welfare-reducing choices when ratings are present.
- This is especially concerning because it is not clear what platforms can do to mitigate this issue.