Title

User-Generated Star Ratings Are Not Inherently Comparable: How Star Ratings Structure Leads to Poor Choices

Permalink

https://escholarship.org/uc/item/65d4t7v3

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 44(44)

Authors

Meister, Matt
Reinholtz, Nicholas

Publication Date

2022

User-Generated Star Ratings Are Not Inherently Comparable

Matt Meister & Nicholas S. Reinholtz

Abstract

User-generated ratings — often elicited and presented as "star ratings" — have become a ubiquitous feature of the online consumer experience. While most research agrees that these user-generated ratings influence individual consumer decisions and overall consumer demand, there is less consensus as to whether user- generated ratings help consumers make better, welfare-enhancing decisions. In this manuscript, we expound on an intrinsic problem with the use of user-generated ratings in product choice decisions. Specifically, product ratings are typically given in an isolated (non-comparative) context, but are typically used in a comparative context, where relative differences in ratings may not reflect relative differences in quality. We provide a simple empirical demonstration of how this structural misalignment can lead consumers to choose suboptimal products and, ultimately, yield reduced consumer welfare.

Keywords: user-generated ratings, online ratings, evaluability

Introduction

User-generated ratings (e.g. star ratings on Amazon.com) have become a ubiquitous feature of the online consumer experience. A recent survey suggests that over 90% of prospective buyers consider user- generated star ratings before making purchase decisions (Qualtrics 2021). Consumers unsurprisingly tend to purchase options with higher star ratings (Chen, Wang & Xie 2011; Chintagunta, Gopinath & Venkataraman 2010; Dellarocas, Zhang & Awad 2007; Hennig-Thurau, Wiertz & Feldhaus 2014). Despite their omnipresence and influence, the degree to which user-generated ratings help consumers make better choices remains an unsettled question. Some argue that ratings are a boon for consumer welfare, and that they allow consumers to enhance their experienced utility (Simonson & Rosen 2014). Meanwhile, others question the diagnostic value of star ratings and suggest they might lead consumers to choose lower quality options (de Langhe, Fernbach & Lichtenstein 2016).

In this manuscript, we describe what we see as an inherent issue with star ratings due to a structural misalignment between their procurement and use. Past purchasers rate products in a non-comparative (isolated) context, but prospective purchasers often use ratings to discern between multiple products. In isolation, raters are likely to focus on aspects which are inherently evaluable (Hsee 1996) or use internal reference points (Birnbaum 1999), such as the alignment of the consumer experience with expectations (Luca & Reshef 2021; Oliver 1980). Thus, ratings provided in an isolated context may be ordinally inconsistent – objectively inferior alternatives can easily be rated higher than objectively superior ones. Because websites present ratings comparatively, they may influence prospective consumers to choose inferior options.

We assess these claims in a series of studies. To create an incentive-compatible paradigm, our "products" are two tasks that differ only in bonus payments — similar to two products that differ only in quality. One task is objectively better than the other (weakly dominant) but receives a significantly lower average rating than the inferior task. When shown these ratings, prospective consumers (future workers) are influenced toward the worse option — in this case, selecting the task with a lower expected bonus payment — even when we provide objective pay information. We provide additional data to cast doubt on alternative accounts of our findings.

Especially troubling is that it is not clear how platforms can mitigate the issue we highlight. Though past work has identified obstacles hindering the curation of beneficial user-generated ratings, they seem surmountable. The prevalence of "fake" reviews creates a concerning degree of uncertainty (Anderson & Simester 2014; Luca & Zervas 2016; Mayzlin, Dover & Chevalier 2014; Stern 2018), but more rigorous standards for posting limit their impact. Issues like small sample size (de Langhe et al. 2016; Powell, Yu, DeWolf & Holyoak 2017), self-selection (Bondi 2019; Li & Hitt 2008), and ulterior motives of raters (Hu, Zhang & Pavlou 2009; Schoenmueller, Netzer & Stahl 2020) can be overcome by encouraging a larger and more representative population to rate. The issue we raise is structural. Platforms can work to deemphasize ratings — or highlight additional information — but, ultimately, our work suggests that consumers must exert greater caution when using star ratings comparatively.

Literature Review

User-Generated Star Ratings

The promise of user-generated ratings is simple and vast. Ratings allow prospective consumers to learn from experiences of prior consumers. As a result, average ratings should represent a simple-to-process summary of normal

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

people's normal consumption experiences. Due to their perceived simplicity and the fact that websites collect ratings on consistent scales (e.g. Amazon's 1-5 star system), star ratings are often used as a point of comparison between products. However, ratings are not inherently comparative.

Ratings are typically produced in isolation after a product is consumed in isolation: The consumer is not prompted to consider or experience alternatives before submitting their rating. Some believe this non- comparative evaluation is a chief benefit of user-generated ratings (Simonson 2016). Presumably, user- generated ratings communicate experienced utility (Simonson & Rosen 2014), indicating what it is like to own product A vs B. If consumers enjoy owning A more than B, they should rate A higher on average. However, this assumes that a star rating conveys experienced utility alone and is not impacted by other aspects of products. In reality, ratings are impacted by things other than experienced utility (e.g. expectations; Luca & Reshef 2021; Oliver 1980), leading star ratings unfit for comparative choice.

Joint-Separate Evaluations

Any rating depends on the rater's mental context at the time of evaluation (Lynch, Chakravarti & Mitra 1991; Parducci 1982), which provides a frame of reference. For example, "How's the weather?" depends on contextual reference points — 40 degrees Fahrenheit is warm for Toronto in January, but unthinkably cold for Santa Fe in July.

In a simple study, Birnbaum (1999) showed that between-subjects judgments (where participants judge one stimulus in isolation) could lead to rating patterns that were both illogical and opposite to what is observed from within-subjects judgments (where participants judge stimuli jointly). Birnbaum asked participants to rate how large either the number 9 or 221 was on a 10-point scale. Despite 9 being objectively smaller than 221, between-subjects ratings suggested the opposite (M9 = 5.13, M221 = 3.10). Birnbaum's explanation is as simple as his experiment. When asked to rate how large a number was, participants had to construct their own frame of reference. For 9, they likely considered a context of single-digit numbers (e.g., 0–9), compared to which 9 seems large. For 221, participants likely considered the context of triple- digit numbers (e.g., 100–999), where 221 is relatively small. Within their respective, isolated frames, 221 is smaller than 9, despite its objective superiority. Though simple, we argue Birnbaum's experiment is less divorced from the reality of star ratings than one might hope. Just because every product on Amazon is rated on a 1–5 scale does not mean that ratings are inherently comparable. Products engender different expectations, which affects how experienced quality is translated on to rating scales (Oliver 1980; Parasuraman, Zeithaml & Berry 1985; 1988).

Consumers are likely aware of this incomparability in some cases (e.g., for products with substantial price differences). Still, we argue that contextual differences arise even for incredibly similar products, as even they can evoke different frames of reference in isolated evaluations. We predict that in isolated evaluation contexts, inferior alternatives may receive higher ratings than their superiors (*Hypothesis 1*). Specifically, this should happen when the better product engenders a less favorable frame of reference than the worse product. For example, if the better product markets itself in such a way that it significantly raises consumers' expectations.

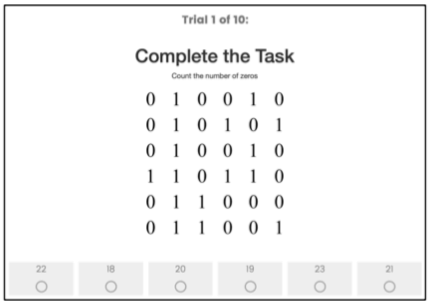
This prediction is problematic because ratings are solicited in isolation, but often used comparatively. Prospective consumers are interested in the comparison raters do not make — how options match up within their choice set. To make this comparison, prospective consumers use information that seems comparable across alternatives (Kivetz & Simonson 2000; Slovic & MacPhillamy 1974). Such is the trouble with star ratings. They seem comparable. They seem as though differences between them reflect meaningful differences between alternatives. The result is that the mere presence of ratings can lead consumers to make objectively poor decisions. Specifically, the presence of user-generated ratings will lead to inferior options being chosen more frequently when they have a higher average star rating than superior options (*Hypothesis 2*).

Empirical Overview

Empirical Stimuli

To test our hypotheses in incentive-compatible experiments, we created two tasks, which we offer to workers on Amazon's Mechanical Turk (AMT) platform. These tasks serve as "products" in our studies. While it may seem strange to call AMT tasks "products", our tasks contain the important features of any consumer choice. They have cost — in our case effort — and provide a benefit — in our case a bonus payment. The tasks also give us tight experimental control: We hold every aspect besides quality constant and avoid issues like fake ratings and self-selection in ratings. Relative differences in objective quality are clearly established as our tasks only differ in bonus payment. We also mitigate concerns about differences in preferences by using professional participants, who consistently value pay rate over other task attributes (Kees, Berry, Burton & Sheehan 2017; Buhrmester, Kwang & Gosling 2016; Paolacci & Chandler 2014). Thus, while our paradigm may lack superficial "mundane realism," it sufficiently captures the important features of the context to provide external validity (Aronson & Carlsmith 1962; Lynch 1982).

In each task, participants are presented a screen of 36 ones and zeros in a 6 × 6 grid, and asked to report the number of zeros in that grid (see Figure 1 for example stimuli; adapted from Abeler, Falk, Goette & Huffman 2011). They repeat this 10 times, with their particular 10 grids drawn randomly from a pregenerated set of 57. By construction, our two tasks require the same amount of effort: the *price* of each product is the same.



Trial 1 of 10:

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

22	18	20	19	23	21
○	○	○	○	○	○

Figure 1. Example stimuli from experimental tasks. The only difference between tasks was the bonus structure, so the stimuli were drawn from the same population of replicates.

Figure 1: Example stimuli for the experimental task.

In both tasks, participants are able to earn a bonus payment determined in part by the number of grids solved correctly. But, within each task, the amount earned per correctly solved grid was determined by chance. Importantly — the possible bonuses participants could earn varied across the two tasks. In both tasks, participants were aware of the possible bonus amounts — and their likelihoods — before undertaking the zero-counting grids, but the actual bonus payment was not determined until after all grids had been completed (see Table 1 for tasks' structures).

Table 1: Structure of Each Task

	Better task	Worse task
Likely Bonus Rate (90%)	5¢	5¢
Unlikely Bonus Rate (10%)	25¢	4¢
Expected Bonus Rate	7¢	4.9¢
Task	Count zeros in grid	Count zeros in grid
Number of Trials	10	10
Size of Grid	6x6	6x6

In both tasks, there was a 90% chance the participant would earn 5¢ per correct answer. In the "better" task, there was a 10% chance one would instead earn 25¢ per correct answer. In the "worse" task, there was a 10% chance one would earn only 4¢ per correct answer. Thus, the "better" task was weakly dominant — its minimum bonus payment was equivalent to the maximum bonus payment in the "worse" task — and was associated with a 43% increase in the expected value of bonus payment (EV_{worse} = 4.9¢/answer, EV_{better} = 7¢/answer). However, the tasks are unlikely to evoke the same frame of reference. In the objectively "better" task, participants' likely payment was the minimum — at the bottom of raters' likely frames of reference (getting 5¢/answer instead of 25¢). In the objectively "worse", one's likely payment was the maximum — the top of these raters' likely frames (getting 5¢/answer instead of 4¢).

The tasks were designed such that people who complete the better task should typically be disappointed in their bonus (5¢ vs. 25¢), whereas people who complete the worse task should typically be relatively pleased (5¢ vs. 4¢). As the bonus payment of our tasks is analogous to the benefits received from a product, we consider this disappointment in pay analogous to a consumer who is disappointed in the quality of a purchase they have made. We note this could occur whenever two products create systematically different expectations (e.g., through advertising, word-of-mouth, pricing, etc.), which is often a marketer's goal.

Our paradigm was designed to maximize internal validity, allowing us to demonstrate the basic problem we highlight. Though constructed to be a strong demonstration, the basic structure is analogous to online marketplaces. In reality, products create systematically different expectations (e.g. through targeting, advertising, etc.). Our tasks do so through their unlikely bonus rates. In reality, ratings are affected by expectations. Our ratings are strongly affected by strong differences in expectations. In reality, prospective consumers think ratings are comparable and informative. Our prospective participants think the same.

Summary of Findings

We began by randomly assigning participants to perform and rate either the better or worse task. As predicted, participants rated the objectively better task significantly lower than the objectively worse task (H1). In phase two of Study 1, (new) participants chose to undertake one of the two tasks. Those who saw objective descriptions of the tasks alone (without star ratings) overwhelmingly selected the better task. However, the mere presence of star ratings led other participants to make worse decisions — choosing the objectively worse task — which we attribute to ratings' illusory comparability (H2).

We then address a potential critique: Maybe participants actually generated higher utility from the task we call "objectively worse". In Study 2, participants undertook both tasks and then chose to repeat one.

We also find that the issues we observe are not easily resolved. In Study 1 and 1A, participants who were shown objective information when selecting a task gave similar star ratings to those who were randomly assigned. This suggests that the effects we observe would not be mitigated by temporal dynamics of the consumption-rating cycle. In Study 3, we find no evidence that presenting text reviews in conjunction with ratings solves the issue we highlight.

Every study was pre-registered and, with the exception of Study 3, incentive-compatible. Data, code, and materials — including pre-registration documents — can be found at https://osf.io/s5fn9/?view_only=044cbda36c3043d39dd9472dd5362ef2).

Studies

Study 1

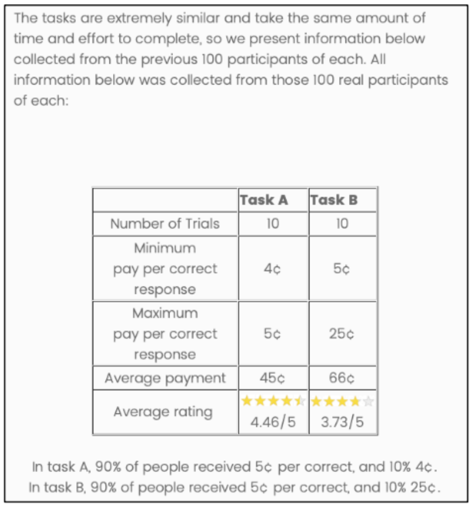
Phase One — Participants & Procedure. We recruited 231 participants from AMT to complete phase one for 50¢ plus whatever bonus payment they earned from completing our task.

We randomly assigned participants to one of our two tasks, which again, only varied in the possible bonuses one could receive. Condition specific bonus information was transparently presented to participants prior to their beginning the task. For example, instructions in the better task stated: “When you finish counting zeros, you will be randomly selected to either receive 25¢ or 5¢ per correct answer. You will specifically have a 10% chance to receive 25¢ per correct answer, and a 90% chance to receive 5¢ per correct answer.” Participants then answered an attention check question, and — as pre-registered — were removed from the study if they failed, leaving 201 participants.3 Next, participants completed 10 trials of the zero-counting task before being shown their score and the possible bonuses they could earn. On the next page they were informed of their actual randomly determined bonus rate and total bonus. Finally, participants were asked to rate the task on a discrete 1–5 star scale, much as they would a product on Amazon.com, and were provided the option to write a review.

Phase One — Results. We predicted that participants in the better task — despite earning more money on average than those in the worse task (total pay including bonus: M_{better} = \$1.14 vs. M_{worse} = \$0.93) — would be disappointed by being paid at the low end of their frame of reference and thus give lower star ratings for the task (H1). Our results support this prediction ($M_{\text{better}} = 3.73$ vs. $M_{\text{worse}} = 4.44$, $F(1, 199) = 22.91$, $p < 0.001$, $d = .68$), and do not meaningfully differ if we include the number of correct responses and its interaction with condition in the analysis.

Phase Two — Participants & Procedure. In phase two, we assessed whether showing new participants the (real) average star ratings from phase one would affect their choice between tasks. We predicted that showing the star ratings in a comparative context would bias people toward choosing the worse task.

We randomly assigned 533 participants to one of three between subject conditions. In each condition, participants were able to select which of two tasks they would complete. The tasks were the same as those used in phase one. We told participants the tasks were “extremely similar and take the same amount of time and effort to complete.” We varied the additional information we gave participants about the two tasks depending on condition. In the Pay condition, participants were shown the minimum, maximum, and average pay per correct answer received. Below that information, we explicitly stated the payment structure in sentence form. In the Stars condition, we only showed participants the average star rating for each task. In the Stars & Pay condition, we combined the information, with the stars presented below the minimum, maximum, and average pay (Figure 2). The Pay and Stars information were designed to be analogous to the type of information a consumer would see about a product online: product specifications (Pay) and user- generated reviews (Stars).



The tasks are extremely similar and take the same amount of time and effort to complete, so we present information below collected from the previous 100 participants of each. All information below was collected from those 100 real participants of each:

	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	★★★★⯪ 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¢.

Figure 2: Stimuli for the “Stars & Pay” information condition, where the objectively better task was A. Stimuli for the Pay condition were identical, with stars removed.

We counterbalanced whether the better task was presented as “Task A” or “Task B”. This did not affect task choice in this or future studies, so we do not discuss it further. After participants selected their task, they followed the same procedure as those in phase one: attention check, 10 grids of ones and zeros, bonus payment draw, and star rating. In accordance with our pre-registration, we removed those who failed the attention check (which did not differ by condition), and those who answered less than five grids correctly. This left us with 497 participants. Our results did not change if we retain those who answered less than five grids correctly.

Phase Two — Results. Because phase one ratings for the objectively worse task were higher than those for the objectively better, our prediction was that those in the Stars condition would select the worse task more frequently than those in the Pay condition. Further, we predicted that those in the Stars & Pay condition would also select the worse task more frequently than those in the Pay, as the mere presence of star ratings would mislead.

We found support for both predictions. Participants in the Stars information condition selected the worse task more frequently than those in the Pay condition (Pay = 4.3%, Stars = 92.3%, $z = 11.58$, $p < .001$). Similarly, albeit less dramatically, those who saw Stars & Pay selected the worse task more frequently than those who saw pay alone (Pay = 4.3%, Stars & Pay = 15.1%, $z = 3.10$, $p = .002$), indicating that mere exposure to star ratings was detrimental, even in the presence of clearly diagnostic information. Adding star ratings to pay information was associated with a ~350% increase in choosing the objectively worse task, and choosing the worse task was associated with a 17% reduction in total compensation (total pay including bonus: M_{better} = \$1.15 vs. M_{worse} = \$0.96).

Discussion. Study 1 illustrates our simple point: Because star ratings are elicited in isolation they can mislead when used to compare alternatives. Participants in the Stars & Pay condition who chose the worse task presumably did so because they thought ratings conveyed information that was otherwise absent (or ignored payment information). This is the result of ratings' illusory comparability.

A potential concern with Study 1 is that phase one's rating procedure may not match real rating procedures. Phase one participants were randomly assigned to their task and did not know of the other task or it's objectively better/worse pay structure. Real consumers often select a product through comparison, considering multiple options before purchase. If they carry this joint evaluation mode into rating, it is possible that real raters do not rate in isolation.

An exploratory analysis suggests this is not the case, but that ratings are a stubbornly isolated evaluation. Considering only the 329 participants in the Pay and Stars & Pay conditions (who knew both payment structures), those who selected the objectively worse task rated it higher than those who selected the better task (M_{better} = 3.91, M_{worse} = 4.78, $F(1, 327) = 18.16$, $p < .001$, $d = .79$). Study 1A (Web Appendix A) replicates this finding with a pre-registered analysis (M_{better} = 4.16, M_{worse} = 4.77, $F(1, 332) = 15.41$, $p < .001$, $d = .64$). This suggests that the results we observe could emerge and persist in a dynamic context more consistent with real online marketplaces.

Study 2

A possible critique of our interpretation of Study 1 is that participants may actually prefer the objectively worse task. Maybe the disappointment that comes from not getting the 25¢ bonus outweighs the fact that the better task always pays at least as well as the worse. If so, the ratings would indeed reflect the experienced utility participants obtained and choosing the higher-rated, but lower paying, task is not actually a mistake. This is unlikely, as AMT respondents strongly value financial compensation (Kees, Berry, Burton & Sheehan 2017; Buhrmester, Kwang & Gosling 2016; Paolacci & Chandler 2014). Nevertheless, we address this concern in Study 2.

Participants & Procedures We recruited 112 participants from AMT to complete both tasks (order and labels counterbalanced). Consistent with our pre-registration, 12 were removed for failing our attention check. The remaining 100 completed the first task, observed their bonus payment, and completed the other task in the same manner. After participants completed both tasks, we gave them the choice to repeat either task. Consistent with our interpretation of Study 1, we predicted that most would select the objectively better task.

Results Results support our prediction. Of the 100, 82 chose to repeat the objectively better task — significantly more than 50% ($\chi2$ = 21.41, p < .001). Participants' choices seem to reflect a belief that the better task yields more experienced utility, despite being rated lower.

Study 3

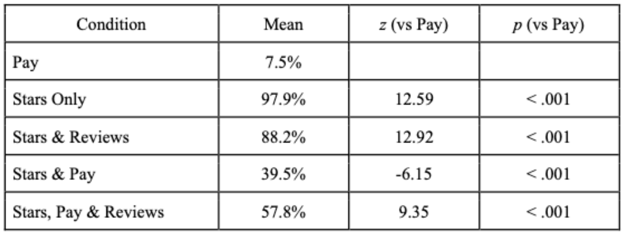
Star ratings impact demand on their own (Luca 2016; Floyd et al. 2014; Rosario, Valck & Sotgiu 2020), but text reviews contain information that stars do not (Tirunillai & Tellis 2014). Potentially, this information could indicate a rater's frame of reference. For example, written reviews for our better task could explain that someone rated the task poorly because they "only" got paid 5¢, while reviews for our worse task could explain someone was happy to earn 5¢. If text reviews contained such information, consumers could discover the reason that our objectively superior task was rated lower.

The inclusion of text reviews is unlikely to solve ratings' problem. Because reviews are also written in isolation — alongside ratings — it is unclear why reviewers would explicitly provide contextual information they don't consider when rating. Such information must be provided incidentally, which is possible but unlikely. We tested this in Study 3, predicting that adding text reviews would not solve ratings' penchant to mislead.

Participants & Procedure 928 participants were recruited from AMT. This study followed a similar design and procedure to phase two of Study 1. However, this study included five conditions — three replicate Study 1, while the two new conditions present text reviews (collected in phase one of Study 1) below either star ratings alone (Stars & Reviews condition) or star ratings & pay information (Stars, Pay & Reviews condition).

Another important difference in Study 3 is that choice in this study was hypothetical. We did not have participants complete the 10 trials or earn a bonus, as we already had incentive compatible evidence from two previous studies. Lastly, we presented additional irrelevant information in each of the "Pay" conditions (Pay, Stars & Pay, and Stars, Pay & Reviews) that did not differ between tasks to each condition (see Web Appendix B for stimuli). This was done to decrease the salience of pay information, which we felt made prior Stars & Pay conditions unrealistically easy. In a sense, those studies present a "lower bound" of the mere influence of stars, while this study approaches possible "upper bounds".

Table 2: Results of Study 3



Condition	Mean	*z* (vs Pay)	*p* (vs Pay)
Pay	7.5%		
Stars Only	97.9%	12.59	< .001
Stars & Reviews	88.2%	12.92	< .001
Stars & Pay	39.5%	-6.15	< .001
Stars, Pay & Reviews	57.8%	9.35	< .001

Results Table 2 contains our replications of H2. Comparing each condition's likelihood to select the worse task to that of the Pay condition suggests in each case that star ratings misled.

To test whether text reviews helped participants make better choices, we analyze the four conditions who see star ratings as a 2 (pay info: yes vs. no) × 2 (text reviews: yes vs. no) between-subjects. This analysis finds that seeing text was marginally detrimental, leading to more frequent selection of the worse task ($M_{TextReviews} = 73.0\%$, $M_{NoReviews} = 68.9\%$, $z = -1.84$, $p = .066$). Adding reviews to the Stars & Pay condition harmed choice quality ($z = -3.46$, $p < .001$). Although the Stars & Reviews condition did make better choices than the Stars condition ($z = -3.29$, $p = .001$), their choices remained significantly worse than chance ($\chi 2 = 61.09$, $p < .001$).

General Discussion

Across three studies we found: 1) An objectively worse task can receive higher average ratings than an objectively better task, and 2) when given the choice between tasks, participants who see those ratings select the objectively worse task more frequently. One may question the degree to which our paradigm maps reality. It lacks superficial, or "mundane" realism (Aronson & Carlsmith 1962), but we believe the important features (price/effort, quality/bonus payment) are present to justify external validity, particularly given the internal validity benefits conferred by our experimental approach (e.g., no selection effects, no fake reviews, etc.). Our results suggest that at any given price point, online marketplaces may be plagued with higher-rated, but lower quality products. We hope future work will examine the degree to which our findings will generalize to the real world.

Despite the potential concerns, our evidence demonstrates a basic point: star ratings are used to make comparisons they do not speak to. They are not a measure of relative quality nor satisfaction and are not inherently comparable. Consumers are implicitly aware of this — we think it unlikely anyone would purchase a pair of $10 headphones instead of a $500 pair solely because the $10 pair was rated higher. But differences in expectations aren't always obvious. One might expect ratings for alternatives at the fringes of different price tiers (i.e. the most expensive budget hotel, cheapest luxury car, etc.) to be most misleading. Our results should be most relevant in categories where consumers' expectations vary widely for alternatives. However, such contexts would exacerbate, not create, the flaw we identify. For example, Luca & Reshef (2021) find that relatively small (3%–9%) increases in restaurants' prices lead to decreased ratings on Yelp.com for those same restaurants, indicating that expectations affect ratings in ways consumers might overlook, naively assuming that differences in ratings reflect meaningful differences between alternatives, not simply differences in expectations.

Our Stars & Pay conditions should have made differences in expectations relatively obvious. Stimuli made it clear that one task contained the possibility of a very high payout, but that that payout was unlikely. Some participants understood this and weren't influenced by star ratings. However, a significant number still selected the worse task. This suggests that users of ratings' default belief is that differences convey differences in alternatives themselves, not in expectations they engender.

There are also some ways for platforms to mitigate ratings' detrimental effects. The simplicity and similarity of our tasks allowed people to understand objective information, which they used to make significantly better decisions when it was presented. Because people strongly weigh information that is comparable across alternatives (Kivetz & Simonson 2000; Slovic & MacPhillamy 1974), we suggest platforms make objective information more easily understandable, comparable, and accessible for between- product comparisons. However, even this suggestion is not simple. In Study 3, the addition of more objective information appeared to harm choice quality, indicating that platforms have to be careful not to overwhelm with irrelevant information.

In proclaiming the benefits of star ratings, Simonson writes about their ability to communicate the "absolute value" of alternatives (Simonson & Rosen 2014; Simonson 2016). Simonson uses one definition of absolute: Total value — not diminished or qualified in any way. But absolute has a second definition: Viewed independently — not relative or comparative. Our data suggest that the second definition fits perfectly. User-generated ratings likely are a valid way to learn about individual alternatives — especially how alternatives met expectations. This use fits their structure. Unfortunately, ratings are not designed to be comparable across alternatives and, as we demonstrate, can easily mislead consumers who use them comparatively.

References

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. *Proceedings of the tenth annual conference of the cognitive science society* (pp. 510-516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill.

Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. *Cognition and Brain Theory*, *6*, 287-317.

Matlock, T. (2001). *How real is fictive motion?* Doctoral dissertation, Psychology Department, University of California, Santa Cruz.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ohlsson, S., & Langley, P. (1985). *Identifying solution paths in cognitive diagnosis* (Tech. Rep. CMU-RI-TR-85-2). Pittsburgh, PA: Carnegie Mellon University, The Robotics Institute.

Shrager, J., & Langley, P. (Eds.) (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.

Aronson, E., & Carlsmith, J. (1962). Performance expectancy as a determinant of actual performance. *The Journal of Abnormal and Social Psychology*, *65*(3), 178.

Abeler J., Falk, A., Goette, L., & Huffman, D. (2011). Reference points and effort provision. *American Economic Review, 101*(2), 470-92.

Anderson, E.T., & Simester, D. (2014), Reviews without a Purchase: Low Ratings, Loyal Customers, and Deception. *Journal of Marketing Research*, 51(3), 249–69. https://journals.sagepub.com/doi/abs/10.1509/jmr.13.0209

Birnbaum, M.H. (1974). Using contextual effects to derive psychophysical scales. *Perception & Psychophysics, 15,* 89-96.

Birnbaum, M.H. (1982). Controversies in psychological measurement. Bernd Wegener, ed. *Social Attitudes and Psychophysical Measurement* (Erlbaum, Hillsdale, NJ), 401-485.

Birnbaum, M.H. (1992). Should contextual effects in human judgment be avoided? [Review of E.C. Poulton, Bias in quantifying judgments]. *Contemporary Psychology, 37*, 21-23.

Birnbaum, M.H. (1999). How to show that 9> 221: Collect judgments in a between-subjects design. *Psychological Methods, 4*(3), 243.

Bondi, T., (2019). Alone, Together: Product Discovery Through Consumer Ratings. Working Paper. https://6a105211-fb5d-4365-be97-db88ba922c3b.filesusr.com/ugd/31323e_ba97d8258f65445f9dc2e91ba38c44e3.pdf

Buhrmester, M., Kwang, T., & Gosling, S.D. (2016). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data?. *Perspectives on Psychological Science, 6*(1), 3–5.

Chen, Y., Wang, Q.I., & Xie J. (2011). Online Social Interactions: A Natural Experiment on Word of Mouth versus Observational Learning. *Journal of Marketing Research*, *48*(2), 238–54.

Chintagunta, P.K., Gopinath, S., & Venkataraman, S., (2010). The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets. *Marketing Science*, *29*(5), 944-957.

Dellarocas, C., Zhang, X.M., & Awad, N.F., (2007). Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures. *Journal of Interactive marketing, 21*(4), 23–45. https://www.sciencedirect.com/science/article/pii/S1094996807700361.

Dellarocas, C. (2003), The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms. *Management Science*, *49*(10), 1407–24.

Greenwald, A.G. (1976). Within-subjects designs: To use or not to use?. *Psychological Bulletin, 83*, 314- 320.

Hennig-Thurau, T., Wiertz, C., & Feldhaus, F. (2014). Does Twitter Matter? The Impact of Microblogging Word of Mouth on Consumers' Adoption of New Movies. *Journal of the Academy of Marketing Science*, *43*(3), 375–94. https://www.academia.edu/download/52789644/Assignment_2_Henning- Thurau_et_al._2014_JAMS.pdf.

Hsee, C.K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational behavior and human decision processes*, *67*(3), 247- 257.

Hu, N., Zhang, J., & Pavlou, P.A. (2009). Overcoming the J-Shaped Distribution of Product Reviews. *Communications of the ACM, 52*(10), 144-147. http://portal.acm.org/citation.cfm?doid=1562764.1562800

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising, 46*(1), 141-155.

Kivetz, R., & Simonson, I. (2000). The effects of incomplete information on consumer choice. *Journal of Marketing Research*, *37*(4), 427-448.

De Langhe, B., Fernbach, P.M., & Lichtenstein, D.R. (2016). Navigating by the Stars: Investigating the Actual and Perceived Validity of Online User Ratings. *Journal of Consumer Research*, *42*, 817–33.

Li, X., & Hitt, L.M. (2008). Self-Selection and Information Role of Online Product Reviews. *Information Systems Research*, *19*(4), 456–74. http://pubsonline.informs.org/doi/10.1287/isre.1070.0154.

Luca, M., & Reshef, O. (2021) The Effect of Price on Firm Reputation. *Management Science, 67*(7), 4408- 4419.

Luca, M., & Zervas, G. (2016). Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. *Management Science*, *62*(12), 3412–27. http://pubsonline.informs.orghttp//www.informs.org.

Lynch, J.G. Jr (1982). On the external validity of experiments in consumer research. *Journal of Consumer Research*, *9*(3), 225-239.

Lynch, J.G. Jr, Chakravarti, D., & Mitra, A. (1991). Contrast effects in consumer judgments: Changes in mental representations or in the anchoring of rating scales?. *Journal of Consumer Research, 18*(3), 284-297.

Mayzlin, D., Dover, Y., & Chevalier, J. (2014). Promotional Reviews: An Empirical Investigation of Online Review Manipulation. *American Economic Review*, *104*(8), 2421–55.

Oliver, R.L. (1980). A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions. *Journal of Marketing Research*, *17*(4), 460–69, https://journals.sagepub.com/doi/full/10.1177/002224378001700405.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current directions in psychological science, 23*(3), 184-188.

Parasuraman, A., Zeithaml, V.A., & Berry, L.L. (1985). A conceptual model of service quality and its implications for future research. *Journal of marketing*, *49*(4), 41-50.

Parasuraman, A., Zeithaml, V.A., & Berry, L.L. (1988). SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality. *Journal of Retailing*, *64*(1), 12–40.

Parducci, A. (1982). Category ratings: Still more contextual effects. *Social attitudes and psychophysical measurement*. 89-105.

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.

Qualtrics Survey Software (2021) Online Review Statistics to Know in 2021. Retrieved November 15, https://www.qualtrics.com/blog/online-review- stats/.

Schoenmueller, V., Netzer, O., & Stahl F. (2020). The Polarity of Online Reviews: Prevalence, Drivers and Implications. *Journal of Marketing Research*, *57*(5), 853–77. https://journals.sagepub.com/doi/abs/10.1177/0022243720941832.

Simonson, I., & Rosen, E. (2014). *Absolute value: What really influences customers in the age of (nearly) perfect information*. (Harper Collins, New York).

Simonson, I. (2016). Imperfect Progress: An Objective Quality Assessment of the Role of User Reviews in Consumer Decision Making, A Commentary on de Langhe, Fernbach, and Lichtenstein. *Journal of Consumer Research*, *42*(6), 840–45. https://academic.oup.com/jcr/article-lookup/doi/10.1093/jcr/ucv091.

Slovic, P., & MacPhillamy, D, (1974). Dimensional commensurability and cue utilization in comparative judgment. *Organizational Behavior and Human Performance*, *11*(2), 172-194.

Stern, J. (2018). Is It Really Five Stars? How to Spot Fake Amazon Reviews. *Wall Street Journal*, (December 20).

Tirunillai S, Tellis GJ (2014). Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. *Journal of Marketing Research*, *51*(4), 463–79. http://journals.sagepub.com/doi/10.1509/jmr.12.0106.