# Order Doesn't Matter:
## Influence of interface elements ordering on user rating behavior

**Emily Hastings** and **Richa Sehgal** and **Jefferson Fu** and **Jose de Oliveira**

University of Illinois, Urbana-Champaign

### Abstract

Crowdsourced ratings and reviews are prevalent in a variety of contexts including e-commerce, design critiques, and peer assessments. An ongoing concern among feedback requesters is how to gather the most useful feedback possible from the crowd. In this work, we investigate whether manipulating certain aspects of the feedback interface (i.e., the presence and position of a numerical rating element in relation to a freeform text field) can affect the quality or content of the feedback produced by the crowd. Crowd workers from Amazon Mechanical Turk provided feedback on website designs in a between-participants experiment with four conditions representing different interface configurations: *rating first, critique first, rating only,* and *critique only*. Our results indicate that while including only a freeform text field led to longer reviews with more negative comments, the order in which workers completed the numerical rating and the written review did not affect the ratings given or the content, length, or quality of the written reviews.

## 1. Introduction

Over the last couple of years, there has been a lot of research conducted through crowdsourcing platforms. Using these platforms, it is relatively easy, quick, and inexpensive to get many anonymous workers to perform a variety of tasks, such as taking surveys and providing feedback for creative works. Many of these tasks include asking workers for some form of rating, be it in numeric form, a written critique, or both. Given subject anonymity and that all of the tasks are performed online, it is difficult to account for truthfulness in ratings by workers. As such, it is necessary to provide parameters and design interfaces which will influence workers to perform their tasks as diligently and honestly as possible.

This paper investigates whether the placement of a numerical rating scale within a crowd-based assessment interface affects the critiques of workers and if altering the presence and location of a numerical rating will lead crowd workers to produce written critiques of significantly different length, content, or quality. Specifically, this work investigates the role of psychological and design issues of effects such as cognitive biases (e.g. framing and anchoring) and user interface elements, and attempts to discover how they can be manipulated to effectively influence users to yield ratings and comments better aligned to the goals of interface designers and researchers. This work aims to determine whether these effects are influenced by interface elements (and to what degree), what kind of biases the influences cause on user input, and the expected impact of these influences on the data collected by the interface in comparison to the original expectations of the interface designers and researchers.

In short, the aim of this study is to find out how to use psychology to compose rating and review interfaces that extract the best possible results from crowd workers. In this case, best possible results mean worker ratings that are minimally biased due on behalf of interferences from the aforementioned effects.



## 2. Related Work

Researchers have recently been looking extensively into getting online crowds of paid workers to provide fast and affordable design feedback on creative works (Xu, Huang, and Bailey 2014). A lot of research is also going into helping the crowd achieve the standards of expert-level feedback (Yuan et al. 2016).

Another area of research on peer assessment deals with addressing scalability issues. It is a rapidly growing technique in online learning to help students evaluate their peers' work, especially as course sizes increase and massive open online courses (MOOCs) become more popular (Kulkarni, Bernstein, and Klemmer 2015). Another work examined several ways of framing peer assessment task goals to obtain better and more consistent output (Hicks et al. 2016).

There are also a number of uses of crowdsourced ratings for e-commerce and content websites. Customer reviews are ubiquitous in today's e-commerce systems for understanding market feedbacks on different commodities. There is a study that combines machine learning and crowdsourcing together for better understanding customer reviews (Wu et al. 2015). Shopping these days includes recommending products, writing comments and rating vendors. This recent phenomenon of social shopping involves more user participation and social interaction and thus requires concepts of crowdsourcing and customer generated feedback. A study by Leitner and Grechenig (Leitner and Grechenig 2008) shows the results of an extended analysis of collaborative shopping networks and demonstrates the development

of a representative interaction model.

In addition, there have been a few psychological studies on people forming opinions on viewing visual content. Such studies helped us design our experimental design. In a study by Lindgaard et al. (Lindgaard et al. 2006), the authors assessed how the order of elements play a role in website design. Participants were made to rate the visual appeal of a web homepage presented for 500 milliseconds. Next, this experiment was repeated using the same stimuli but they were shown for only 50 milliseconds. It was observed that people rate websites similarly whether they see them for 500 milliseconds or 50 milliseconds. It is clear from these studies that first impressions form quickly and are consistent (Lindgaard et al. 2006). In another relevant study, different numerical rating scales were analyzed. The data showed no overall statistical differences between the different scales (Lindgaard et al. 2006).

As seen in this brief literature review, while there exist numerous studies on crowdsourced critiques, ratings and peer assessment, there has been limited research on the effect the layout of the interface has on the quality and content of the reviews (Hicks et al. 2016). Thus, this paper aims to study and improve upon rating interfaces to generate better and more useful feedback. This idea immediately follows the lead of previous research done by Hicks et al.'s 2016 CHI paper *"Framing Feedback: Choosing Review Environment Features that Support High Quality Peer Assessment"* on how cognitive biases, such as framing and anchoring effects, influence peer evaluation (Hicks et al. 2016). The study investigated how changes to rubrics, task structure, and work representation (changes in framing) impacted the quality, number of explanations and depth of feedback given by reviewers. However, our study runs deeper on the psychological and design issues of such effects, and tries to discover how they can be manipulated to effectively influence users to yield ratings and comments better aligned to the goals of interface designers.

As such, our work aims to determine whether these effects are influenced by interface elements (and to what degree), what kind of biases will influence user input, and how the expected impact of these influences on the data collected by the interface compare to the original expectations of the interface designer. By increasing the academic community's understanding of this aspect of rating interfaces, we hope to make the knowledge gleaned from these reviews more useful to all involved. If a significant difference in the generated critiques is identified, interface designers can modify their interfaces to produce the sort of feedback most useful to them.

## 3. Research Questions

Our experiment answers the following research questions:

**RQ1**: How does altering the presence and order of numerical rating and freeform critique elements in a review interface affect the numerical ratings left by crowd workers?

**RQ2**: How does altering the presence and order of numerical rating and freeform critique elements in a review interface affect the length, content, and quality of written critiques produced by crowd workers?

Answering these questions will help feedback requesters in a variety of contexts design their interfaces to receive crowd feedback that best meets their needs.

## 4. Method

To answer our research questions, we conducted a between-participants experiment with one factor, the *order* in which the different elements of the review interface appeared. There were four conditions, including two control conditions (*rating only* and *critique only*), where participants only gave one kind of feedback, and two experimental conditions (*rating first* and *critique first*), where participants gave both kinds of feedback in varying orders.

### Participants

We recruited 119 participants from Amazon Mechanical Turk (AMT) to take part in this experiment. To qualify for our tasks, workers had to have at least a 90% lifetime approval rating. Participants earned a base reward of $0.25, and were awarded an additional bonus of $0.75 if their responses indicated that they had put forth a reasonable effort in completing the task. Since the data logged by our interface indicates that most participants took about five minutes to provide feedback, workers earned roughly $12 per hour, which is higher than minimum wage in the United States ($7.25).

The number of participants recruited for each condition, and the demographic breakdown of the sample population is shown in Tables 1 through 3.

| | Rating first | Critique first | Rating only | Critique only | *Total* |
|---|---|---|---|---|---|
| Male | 21 | 15 | 20 | 18 | *74* |
| Female | 7 | 14 | 14 | 10 | *45* |
| *Total* | *28* | *29* | *34* | *28* | *119* |

Table 1: Breakdown by gender x condition

| | Rating first | Critique first | Rating only | Critique only | *Total* |
|---|---|---|---|---|---|
| US | 14 | 12 | 20 | 11 | *57* |
| India | 14 | 16 | 11 | 15 | *56* |
| UK/ CAN | 0 | 0 | 1 | 0 | *1* |
| N/D | 0 | 1 | 2 | 2 | *5* |
| *Total* | *28* | *29* | *34* | *28* | *119* |

Table 2: Breakdown by country x condition

### Procedure

We posted four Human Intelligence Tasks (HITs) on AMT, one for each condition. These were identical apart from differing survey links. The links directed workers to a page on

| | 1 | 2 | 3 | 4 | *Total* |
|---|---|---|---|---|---|
| 18-24 | 4 | 3 | 5 | 5 | *17* |
| 25-34 | 17 | 16 | 18 | 13 | *64* |
| 35-44 | 6 | 5 | 8 | 7 | *26* |
| 45-54 | 0 | 4 | 1 | 2 | *7* |
| 55-64 | 0 | 4 | 1 | 2 | *4* |
| 65 or older | 2 | 0 | 2 | 1 | *5* |
| *Total* | *28* | *29* | *34* | *28* | *119* |

Table 3: Breakdown by age x condition

our website where they first entered demographic information and read the instructions for the feedback task, which was broken into three phases. Once finished with the instructions and demographic section, workers pressed a button to progress to the first phase of the feedback task. Prior to each rating phase, the subject would be presented an intermediary page and a five seconds countdown. This pause would serve both to allow subject to be prepared for ratings, and as cooldown period between phases one and two.

In the first phase, participants were shown a series of 20 screenshots of websites in random order and asked to rate the quality of their design by clicking on a series of hypertext links representing a 9-point Likert item (1="low", 9="high"). Each image was shown for 500 ms, in order to capture workers' first impressions of the designs and provide a baseline to which we could compare their later ratings. Lindgaard et al. found that this is enough time to form a consistent opinion of visual appeal (Lindgaard et al. 2006). We chose a 9-point scale for this rating so that we could directly compare our results to those of Lindgaard; in addition, scales of different lengths have been shown to produce similar responses (Huynh-Thu et al. 2011)(Matell and Jacoby 1971). Websites of both low and high design quality were selected by members of the research team and came from a variety of contexts.

In the second phase of the task, participants were shown two randomly selected websites from the first phase again and had as much time as they desired to give feedback according to their condition. For the control conditions, participants either gave only a numerical rating or only a freeform critique addressing the prompt: "Please provide feedback on this design". Participants in the experimental conditions either gave a numerical rating first, then wrote a critique, or vice versa. In an early run, users were allowed to navigate the rating form interface as they saw fit, and their navigation was recorded for analysis. However, by looking at the behavior of users on early results, we realized that most of the participants of the *critique first* condition (12 out of 18 individuals) opted to give ratings before writing their critique, effectively transforming the experiment in the *rating first* condition. The interface was then modified in order to require users to fill out rating and feedback in the specified order, depending on the condition, eliminating the possibility of navigation in the interface. Screenshots of the final version of the interface are shown in Figures 1 through 7.

The third and final phase of the task asked workers to describe how useful they felt their feedback would be to the designers of the websites shown. We included this question as an attention check; the relevance of the answer was a large factor in whether the worker earned the bonus.

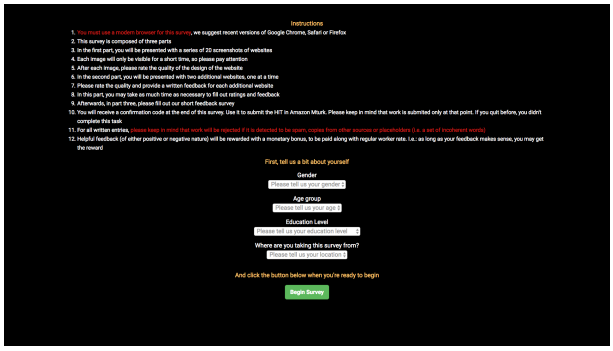


Figure 1: Demographics form

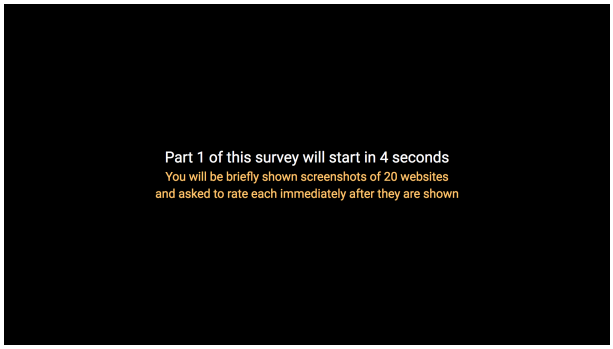


Figure 2: Pause before phase one

After workers submitted their responses to the feedback task on our website, they received a unique, randomly generated code to enter into AMT and submitted the HIT.

## Measures

To answer RQ1, our primary measure was the deviation between a participant's initial rating of a website during the first phase of the feedback task and their subsequent rating during the next phase. Individual ratings were normalized into z-scores, using mean and standard deviations for all ratings given to each individual website in both phases. By using this transformation, similar to that used by Lindgaard et al. (Lindgaard et al. 2006), we avoided differences in scale between the ratings of distincts websites due to notable differences in quality between our chosen assets, making it possible to compare the relative deviation of ratings between phases for all individuals. Thus, our statistical analysis was made over the difference between the z-score of the participant's initial rating and the z-score of the participant's final rating (i.e. rating given on the second phase).

To answer RQ2, we used the same numerical coding schema as Hicks et al. (Hicks et al. 2016). Each piece of written feedback from the second phase of the task was coded by two independent members of the research team

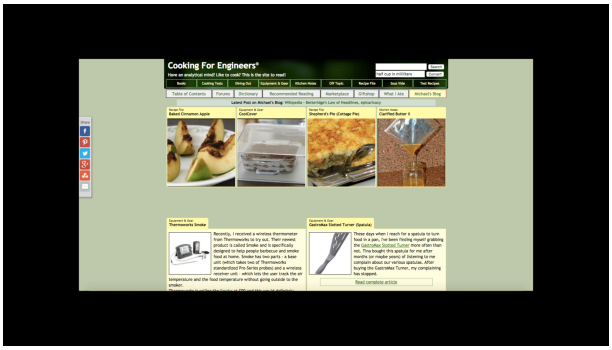


Figure 3: A screenshot for 500ms

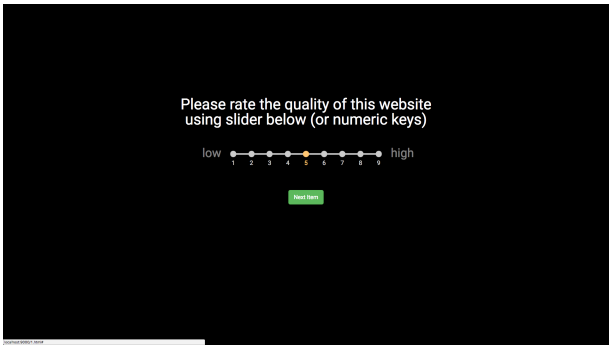


Figure 4: Rating on phase one

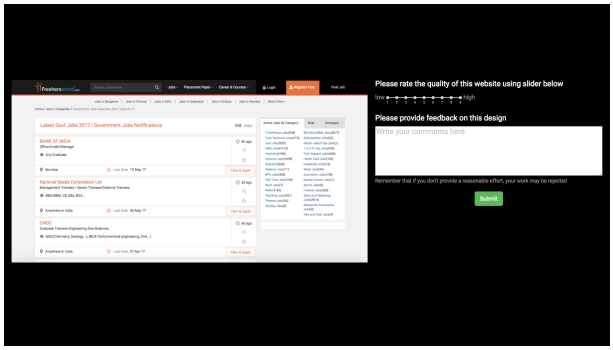


Figure 5: Rating on phase two

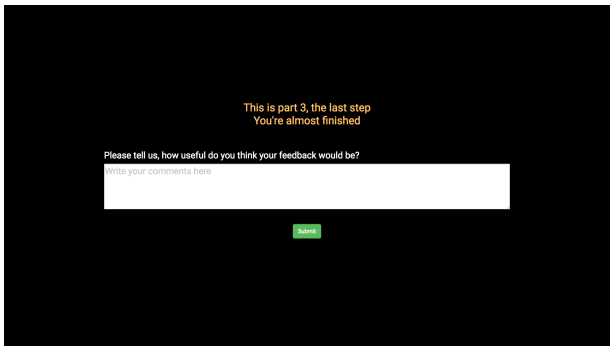


Figure 6: Survey feedback

who were blind to the condition. For each piece of feedback, the raters recorded the number of suggestions with related explanations, the number of positive and negative comments, the word count, and a subjective rating of the quality and helpfulness of the feedback on a scale from 1 (off-topic or unhelpful) to 5 (very helpful). Ratings from the two coders were positively correlated; correlation coefficients ranged from 0.61 for the number of negative comments to 0.88 for the quality rating. All correlations were significant ($p = 0.00$). We therefore used the averages of the coders' judgements for each category in our analysis.

## 5. Results

When analyzing our data, we treated each rating or piece of feedback as a separate data point before they were submitted for ANOVA analysis. Therefore, since each participant reviewed two websites, our sample size for analyzing ratings and critiques, as shown in table 4, was double the number of participants in each condition.

| | Sample size |
|---|---|
| Rating first | 56 |
| Critique first | 58 |
| Rating only | 68 |
| Total | 238 |

Table 4: Sample sizes for rating analysis

### Effects on Ratings (RQ1)

To answer RQ1, we analyzed the deviation of individual scores given in each phase for the *rating first, critique first,* and *rating only* conditions, i.e. the difference between the normalized rating of the second phase and the normalized rating that the individual gave to the same website on the first phase under those conditions. Effectively, due to the normalization procedure, we measured how much individuals deviated from their original rating measured as the proportion of the standard deviation of all ratings for the same website.

**Ratings did not differ significantly between conditions**

An ANOVA did not detect any significant differences between deviations of ratings under conditions ($F(2,180) = 0.86$, $p = 0.43$). We also tested for the absolute value of deviations, with similar negative results ($F(2,179) = 0.82$, $p = 0.44$). A post-hoc Tukey HSD test confirmed that there is no difference between the means of deviations under each condition ($R2 = 0.9076214$). The relationship between descriptive statistics of both deviations and absolute values of deviations may be seen in graphs pictured in Figures 8 and 9.

It is also noteworthy that t-tests on mean deviations failed to confirm the hypothesis that any of them was equal to zero: $p = 0.61$ for *rating first*, $p = 0.01$ for *critique first* and $p = 0.10$ for *rating only*. Although the 95% confidence interval contains zero for both the *rating first* and *rating only* conditions, these results, combined with the fact that all

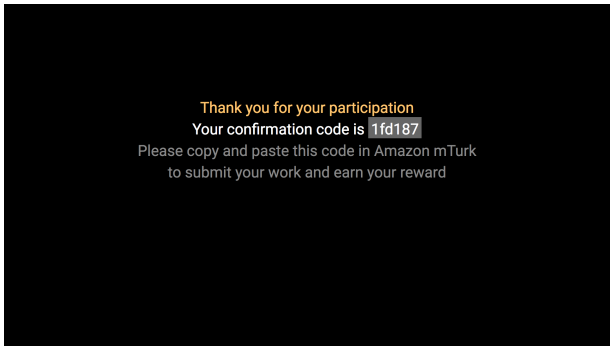


Figure 7: Exit screen

mean deviations were marginally above zero (respectively: = 0.06, = 0.29, = 0.21), indicate that participants generally increased their ratings in the second phase.

Furthermore, regression analysis showed that the deviations are negatively correlated to ratings in the first phase and positively correlated to those in the second phase, although in distinct confidence levels across conditions. In the *rating only* condition, this relationship is strong for the first phase (t = -6.09, p = 0.00); it is slightly less so for the second phase (t = 3.39 p = 0.00). The significance of this relationship is reverted for both the *rating first* condition (t = -2.96, p = 0.01 for the first phase; t = 5.02 p = 0.00 for the second phase), and similar for the *critique first* condition (t = -3.45, p = 0.00 for the first phase; t = 2.59, p = 0.01), although at a lower level. These results indicate that participants tended to deviate more when they gave extreme ratings. A very high rating in the first phase would be followed by a very low rating in the second phase and vice-versa; whereas a median rating would be followed by a closer rating in the second phase. Moreover, this behavior was more accentuated for the *rating only* condition participants than for others. Overall, these results disagree with those from Lindgard et al. (Lindgaard et al. 2006), which observed neglectable deviations between phases on comparable settings (*rating only* condition).

For the remaining control variables, regression analysis showed significant relationship for the *rating only* condition between rating deviations and both number of words written in the final feedback section ($F(1,66) = 6.38$, $p = 0.01$) and time spent in the survey ($F(1,66) = 3.62$, $p = 0.06$), meaning that subjects that spent more time writing feedbacks deviated more from their first phase ratings.

## Effects on Critiques (RQ2)

To answer RQ2, we analyzed the data collected from participants who left written feedback on the website designs (i.e., the *rating first, critique first,* and *critique only* conditions). Average word counts, quality ratings, and numbers of suggestions and positive and negative comments for each condition are shown in Figure 10 to 11.

**Non-numeric critiques contained more negative comments**

An ANOVA did not show any significant differences in the

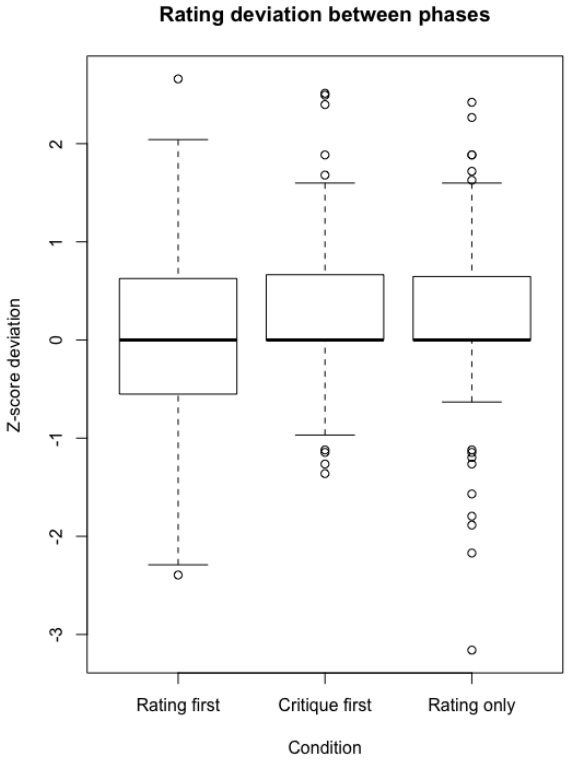


Figure 8: Rating deviation

number of positive comments ($F(2,177) = 0.06$, $p = 0.94$) or suggestions with explanations ($F(2,177) = 0.88$, $p = 0.42$) in the critiques produced by workers in the different conditions. We did, however, observe a significant difference in the number of negative comments ($F(2,177) = 2.97$, $p = .05$). A post-hoc Tukey test revealed that participants in the *critique only* condition produced significantly more negative comments than those in the *rating first* condition ($p = 0.02$), though the effect size is small (Cohen's $d = 0.45$). No other pairwise differences were significant.

These results are contrary to those reported by Hicks et al., who observed no differences in negative comments and found that participants who provided a numeric rating along with a freeform critique gave more suggestions and more positive comments than those who only gave non-numeric feedback (Hicks et al. 2016). Hicks speculated that reviewers providing both a numeric rating and written feedback felt it necessary to console reviewees for their rating, and so wrote with a more positive tone overall. It is possible that a similar phenomenon occurred in our experiment. Non-numeric reviewers may have felt free to write more negatively than numeric reviewers, since they did not have to soften the blow of a low numeric rating.

Alternatively, since participants in this condition did not have a numeric rating to help convey their opinions, these reviewers were only able to express their dislike of the designs through the content of their critiques. Therefore, neg-

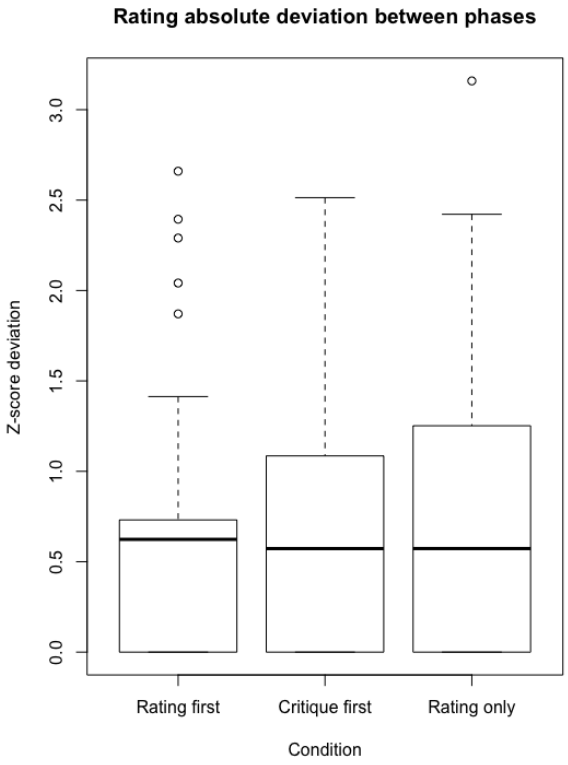


Figure 9: Rating deviation absolute values

ative opinions that might otherwise have been reflected in lower ratings surface as more critical written feedback.

**Non-numeric critiques were longer**

While an ANOVA did not find a significant difference in the quality of the critiques according to condition ($F(2,177) = 1.18$, $p = 0.31$), we did observe a significant difference in the word counts of the critiques ($F(2,177 = 3.795$, $p = 0.02$). A post-hoc Tukey test revealed that the *critique only* condition produced reviews that were on average 13 words longer than the *rating first* ($p = .02$) condition. In addition, though the difference was not significant, the *critique only* critiques were on average ten words longer than the *critique first* condition ($p = 0.11$).

These findings partially agree with those of Hicks et al., who observed that reviewers who provided only non-numeric feedback wrote longer and higher-quality critiques (Hicks et al. 2016). The observation that the *critique only* condition produced longer reviews is not unduly surprising; a likely explanation is that workers were willing or able to designate only a certain amount of time and effort to completing our task, and so those who did not have to perform the additional step of numerically rating designs in the second phase could devote more energy to the written feedback.

It was surprising, however, that there were no significant differences between the *rating first* and *critique first* conditions for any measure. We return to this point in the Discussion.

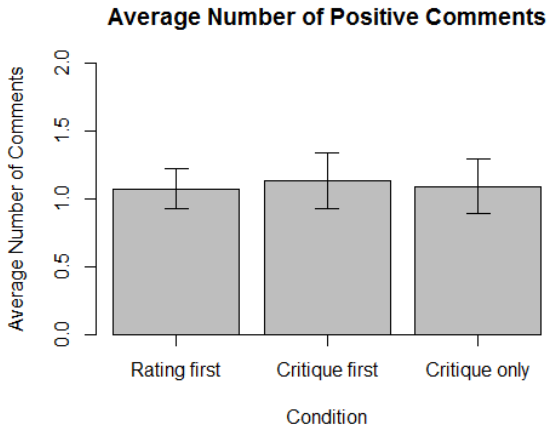


Figure 10: Positive comments

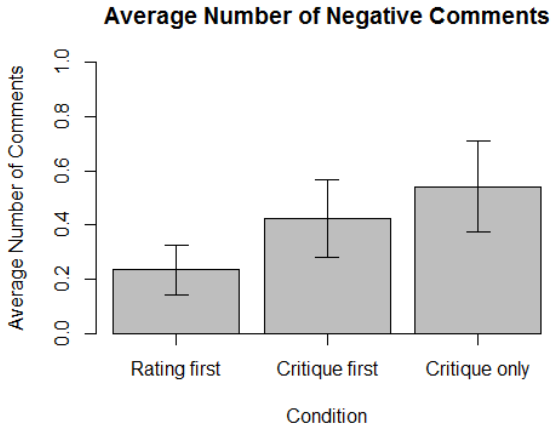


Figure 11: Negative comments

## 6. Discussion

We conducted an experiment illustrating how altering the presence and order of numerical rating and freeform critique elements in a review interface affects numerical ratings rated by crowd workers as well as the length, content, and quality of the written critiques.

The statistical analysis in regards to numerical ratings indicate that there is no discernible difference on the obtained average rating by providing the user with a rating interface element before or after an open feedback interface element. Moreover, no statistical difference was detected whether the user provided feedback or not (i.e. *rating only* condition).

By examining different interface configurations, we also detected that the positioning of elements does not always affect user behavior; when giving users the freedom to navigate interface elements, they generally opted for rating before writing. One possible explanation for this behavior is that, in this setting, users prefer performing the task that de-

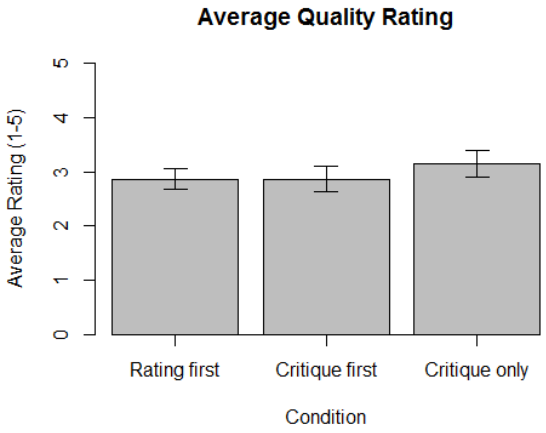


Figure 12: Rating quality

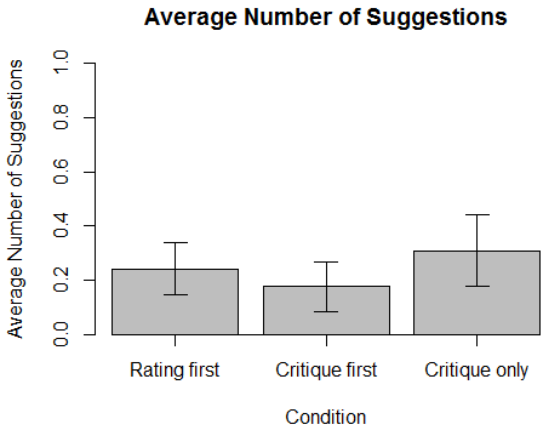


Figure 13: Suggestions

mands lower cognitive effort (clicking on a rating) before the one with higher cognitive effort (writing text). This rationale is inline with the theory of cognitive easiness (Kahneman 2011); it it less costly for the user to repeat a task that they have performed several times (rating) than to engage in a new, complex task. According to Kahneman, the cost comes from demanding the user's brain to switch from System 1 (automatic, subconscious) to System 2 (logical, conscious), which meets automatic resistance.

Given this behavioral-based drawback, the experiment was drawn to an alternate setting by forcing the necessary conditions and removing the freedom of navigation. Nevertheless, guiding users in the desired sequence did not affect results. Once again, there was no detectable difference in the average ratings users provided before or after writing critiques. However, we observed a clear trend in user behavior: subjects who gave extreme (either very high or very low) ratings in the first phase generally deviated more than those

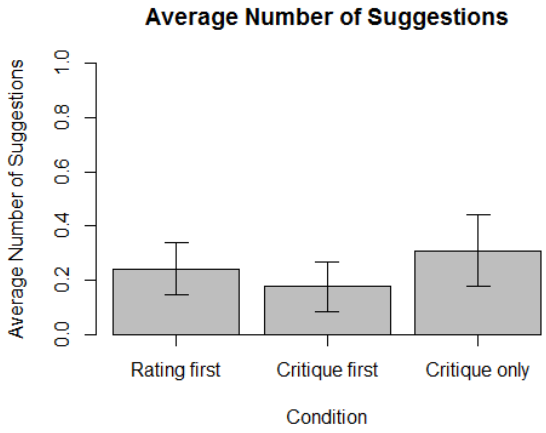


Figure 14: Word count

who gave more moderate ratings. By looking at comments, we detected that participants were aware that, in the second phase, they were rating a website they had already rated in the first phase. It is reasonable to conclude that, by means of recency (Kim and Fesenmaier 2008), those who opted for moderate ratings in the first phase chose a rating for the second phase close to the one recently given for that same asset, and their feedback was written in order to justify his previous choice. Data for the extreme rating behavior hints it likely individuals changed their own rating scales between phases, making it hard to compare deviations across individuals. However, it does not leave room for any conclusions at this point.

While our results indicated that workers who rated using non-numeric critiques left more negative comments and provided a lengthier explanation than their counterparts, there was no significant difference between the either the critiques or the ratings produced by the two experimental conditions (*rating first* and *critique first*). This is unexpected, as we had originally hypothesized that participants in the *critique first* condition might, through a sort of "auto-framing" effect, influence their ratings with the content of the reviews, or else be less likely to simply explain their numerical rating in their critique. A possible explanation for this lack of difference is that participants in the experimental conditions were told in the task instructions they would be producing both a numerical rating and a critique. Therefore, the workers in the *critique first* condition may have already considered the rating they planned to leave as they were writing their feedback, making the two conditions virtually the same.

This study presents several implications for feedback requesters and interface designers. First, for contexts in which more critical written feedback is desired, our results suggest that it is best to include only a free-form critique with no associated numerical rating, since reviews in this format were lengthier and contained more critical, negative comments. Second, if a numerical rating is desired in addition to a written critique, the order of the rating interface (before or after

the free-form critique) is immaterial; its placement did not impact the rating given, nor the content, length, or quality of the critique. In addition, availability bias, which is the tendency to make judgments on the basis of what can be easily be brought to mind (e.g. a recent occurrence) (Tversky & Kahneman, 1974) (Tversky and Kahneman 1975) when evaluating a topic, such as website design, may be important to consider when designing interfaces. Even though the first phase of the task went by very quickly, several workers during the second phase commented that they remembered seeing one of the websites from the first phase as well as their rating for it.

## 7. Limitations

One potential confound to our study is that since participants reviewed the same images in both the first and second phases of the feedback task, they may have remembered their original ratings and been anchored by them when providing their second rating. As discussed, several participants alluded to this phenomenon in their comments. A possible solution to this problem could be to include a distracter task in between feedback phases, in order to increase cognitive load and encourage participants to forget their earlier ratings (Purchase 2012).

In addition, the extreme deviation behavior, in its place, may be attenuated by using a Likert scale on a smaller range. Since previous studies found that distinct scale ranges do not significantly affect statistical results, using a smaller range may help keep deviations constrained to a smaller interval for all participants by incentivizing them to use a smaller number of middle range rating points (Matell and Jacoby 1971), and make deviations more readily comparable across individuals.

Another possible confound is that since participants in the experimental conditions knew that they would be providing both a critique and a numerical rating, as well as the scale of the rating, there was nothing to prevent them from considering their rating before writing, even if they were in the *critique first* condition. Future work could tease apart the effects of this foreknowledge.

## 8. Conclusion

A major concern in the context of crowdsourced ratings and reviews is designing feedback interfaces so as to receive the most useful feedback possible from the crowd. In this work, we investigated whether manipulating the order and presence of numerical rating elements and freeform text fields impacted the value of the ratings or the content, length, or quality of the critiques generated. We conducted an experiment in which workers from Amazon Mechanical Turk provided feedback on website designs using different interface configurations. Our findings suggest that having only a written feedback aspect in the interface prompts users to produce longer reviews containing more negative comments. However, when both a numerical rating and a text component are present, we observed no differences in either the ratings or critiques received, regardless of the order in which the two elements are completed. The implications of this study for feedback requesters are that if longer, more critical feedback is desired, interfaces should include only a written component, and that if both a numeric rating and a written critique are necessary, the order in which they are presented is inconsequential.

## Acknowledgements

We would like to thank the workers on Mechanical Turk who participated in our study, without whom this experiment would not have been possible. We would also like to thank Prof. Brian Bailey, Grace Yen, and our classmates for their valuable feedback on this project.

## References

Hicks, C. M.; Pandey, V.; Fraser, C. A.; and Klemmer, S. 2016. Framing feedback: Choosing review environment features that support high quality peer assessment. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 458–469. ACM.

Huynh-Thu, Q.; Garcia, M.-N.; Speranza, F.; Corriveau, P.; and Raake, A. 2011. Study of rating scales for subjective quality assessment of high-definition video. *IEEE Transactions on Broadcasting* 57(1):1–14.

Kahneman, D. 2011. *Thinking, fast and slow*. Macmillan.

Kim, H., and Fesenmaier, D. R. 2008. Persuasive design of destination web sites: An analysis of first impression. *Journal of Travel research* 47(1):3–13.

Kulkarni, C. E.; Bernstein, M. S.; and Klemmer, S. R. 2015. Peerstudio: rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, 75–84. ACM.

Leitner, P., and Grechenig, T. 2008. Collaborative shopping networks: Sharing the wisdom of crowds in e-commerce environments. *BLED 2008 Proceedings* 21.

Lindgaard, G.; Fernandes, G.; Dudek, C.; and Brown, J. 2006. Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour & information technology* 25(2):115–126.

Matell, M. S., and Jacoby, J. 1971. Is there an optimal number of alternatives for likert scale items? study i: Reliability and validity. *Educational and psychological measurement* 31(3):657–674.

Purchase, H. C. 2012. *Experimental human-computer interaction: a practical guide with visual examples*. Cambridge University Press.

Tversky, A., and Kahneman, D. 1975. Judgment under uncertainty: Heuristics and biases. In *Utility, probability, and human decision making*. Springer. 141–162.

Wu, H.; Sun, H.; Fang, Y.; Hu, K.; Xie, Y.; Song, Y.; and Liu, X. 2015. Combining machine learning and crowdsourcing for better understanding commodity reviews. In *AAAI*, 4220–4221.

Xu, A.; Huang, S.-W.; and Bailey, B. 2014. Voyant: generating structured feedback on visual designs using a crowd of non-experts. In *Proceedings of the 17th ACM conference on*

*Computer supported cooperative work & social computing*, 1433–1444. ACM.

Yuan, A.; Luther, K.; Krause, M.; Vennix, S. I.; Dow, S. P.; and Hartmann, B. 2016. Almost an expert: The effects of rubrics and expertise on perceived value of crowdsourced design critiques. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1005–1017. ACM.