The Design Space of Opinion Measurement Interfaces: Exploring Recall Support for Rating and Ranking

Syavash Nobarany, Louise Oram, Vasanth Kumar Rajendran, Chi-Hsiang Chen, Joanna McGrenere, Tamara Munzner
Department of Computer Science
University of British Columbia, Vancouver, Canada
{nobarany, olouise, vasanthr}@cs.ubc.ca, devswing@interchange.ubc.ca, {joanna, tmm}@cs.ubc.ca

ABSTRACT
Rating interfaces are widely used on the Internet to elicit people's opinions. Little is known, however, about the effectiveness of these interfaces and their design space is relatively unexplored. We provide a taxonomy for the design space by identifying two axes: Measurement Scale for absolute rating vs. relative ranking, and Recall Support for the amount of information provided about previously recorded opinions. We present an exploration of the design space through iterative prototyping of three alternative interfaces and their evaluation. Among many findings, the study showed that users do take advantage of recall support in interfaces, preferring those that provide it. Moreover, we found that designing ranking systems is challenging; there may be a mismatch between a ranking interface that forces people to specify a total ordering for a set of items, and their mental model that some items are not directly comparable to each other.

Author Keywords: rating; ranking; opinion; attitude; judgment; review; design space;

ACM Classification: H.5.2 [Information Interfaces and Presentation]: User Interfaces - Graphical user interfaces.

INTRODUCTION
Rating products and services on the Internet is pervasive. The most common interface allows people to assign a rating on a sequential scale using stars or other indicators; other mechanisms include the diverging scale of like vs. dislike, or the categorical tag of like (Fig. 1). Despite their prevalence, the design space of possible interfaces for measuring subjective opinions has received little attention in HCI research. Work to date has mostly focused on the parameters and visual design of *n*-point Likert scale interfaces (e.g. [1,21,23]), or tailored application-specific designs that, while being more expressive, require significant investment for reuse in other application domains (e.g. [12]). Customer attitudes towards products and services play an important role in communication

Figure 1. Rating interfaces in common use today.

between customers and customer relationship management. It has been shown that presenting user ratings can influence other users' perception of products and services and play an important role in their decisions [7,25].

While studies of opinion measurement interfaces have been very limited in the HCI literature, social psychologists and marketing researchers have long been interested in analyzing methods for eliciting and understanding people's attitudes [14]. The design space of computer interfaces for opinion measurement is relatively unexplored and designers often attend to the aesthetics and number of levels, rather than deeper factors that can affect mental models. We felt a need for a deeper understanding of the design space and started with a formative study to expand our understanding of how and why people rate things on the Internet. Based on findings from that study as well as the literature, we characterized the design space along two core dimensions: Measurement Scale (rating vs. ranking), and Recall Support, which is the information shown to aid opinion formation. To explore this design space, we iteratively designed three alternative interfaces that varied along the dimensions. The interfaces incorporated different approaches to judging, a general term that includes both rating (assigning absolute values) and ranking (relative evaluation by placing items in order). Finally, we conducted a mixed-methods study to assess the interfaces.

The specific contributions of this work are as follows. First we provide a high-level taxonomy for the design space of computer interfaces to measure subjective opinions. Second, we iteratively designed and evaluated several interface alternatives for eliciting opinions within the design space. Third, we ran, to our knowledge, the first controlled experiment that systematically investigates computer interfaces that measure subjective opinions. That experiment has several key findings: (a) that users do take advantage of recall support, and prefer interfaces that provide it over those that do not; (b) that according to self-reports, users generally care more about the ability to be accurate over the ability to be fast, especially if provided with recall support; and (c) that designing ranking systems is challenging, in that there may be a mismatch between a ranking interface that forces people to specify a total ordering for a set of items, and their mental model that some items are not directly comparable to each other. Based on these findings, we conclude that providing recall support without imposing comparison can facilitate judgment, and we present several key directions for future research.

ATTITUDE MEASUREMENT

Although there is no agreement on what an attitude is [9,14], according to one of the often-cited definitions attitudes are "tendencies to evaluate an entity with some degree of favor or disfavor, ordinarily expressed in cognitive, affective, and behavioral responses" [8]. Thurstone differentiated attitude from opinion, by defining opinion as an expression of attitude [22]; however, for the purpose of this study, we found the distinction unnecessary and the two words are used interchangeably.

One approach to designing interfaces for expressing opinion is through understanding the cognitive process that is used to generate opinions. Reporting attitudes has been described as a three-stage process [14]: the first stage is the *automatic activation phase*, in which an initial opinion is formed without an intention or any effort. The second phase is the *deliberation phase*, where relevant information is retrieved from memory. Then ultimately, in the *response phase*, the output of the deliberation and automatic activation phases are turned into a response.

Current online reviewing systems (Fig. 1) use n-point scales to capture and present users' attitudes, and $n > 2$ is the most common practice. For example, systems such as Yelp and Amazon use 5-point ratings. Often n is an odd number and the scale is diverging, so that the mid-point represents neutral and higher and lower positions represent positive and negative attitudes. There are also systems that only allow for positive attitudes using sequential scales such as Michelin guide's 3-point and Facebook's 1-point 'Like' scale. Another common practice is using 2-point diverging scales such as thumbs up/down on YouTube. What all of these systems have in common is an absolute judgment scheme, where each item is rated in isolation from any other items. This allows for a quick and easy way of expressing opinions and displaying aggregated results. For a literature review on rating scales in surveys see [15].

The relative merits of ranking and rating mechanisms for measuring people's attitudes has long been a subject of debate [2,13] and depending on the data being collected, one of the two methods may be more suitable. Some researchers argue that ranking techniques better match the conception of attitudes that are considered inherently comparative and competitive. For example, if one of the important goals of the judgment is choosing an option, ranking can be preferable; however, if the goal is to categorize a set of items, rating can be more appropriate.

Rankings are often more cognitively demanding and require concentration, which is problematic when dealing with a long list of items [2]. The prevalence of using ratings instead of rankings has been mainly to reduce phone survey completion time; however, making the task easier may reduce precision [2]. Moreover, lack of consistency is a known issue of rating systems [4,7] and several mechanisms such as re-rating [3] and bias-from-mean adjustment [11] have been suggested to alleviate the problems of intra-rater and inter-rater consistency.

FORMATIVE STUDY

To our knowledge, there is little empirical evidence as to why and how people rate products and services on the Internet. Ozakca and Lim [17] showed that people tend to give feedback when they have strong opinions. Harper et al. conducted a survey on frequent movie raters that used Movielens [10] to identify motivations for rating movies. They found that improving recommendations, fun of rating, and keeping a list of watched-movies are the most salient motivators. In contrast, we focused on collecting more qualitative data through interview and observation, especially from infrequent raters.

We first interviewed participants about their previous experience with rating items on the Internet (i.e. what, when, how, and using which systems do they rate items on the Internet), as well as their motivation for rating and consuming others' ratings. The interviews included both Likert-scale and open-ended questions. In part two of the study, we asked the participants to rate at least two items from different domains including movies, restaurants, music, recipes and products using various reviewing systems (e.g., IMDB, Yelp). We used a think-aloud protocol, and probed if further explanation of the reasoning behind their rating was needed. The interviews took 25-45 minutes, and were recorded through note taking.

Seven participants (4 females) with various levels of rating experience participated: one was a frequent rater and the rest had occasionally rated items such as recipes, music, and movies. When asked to show the rating systems they used, all participants showed us standard star interfaces.

Results

Part 1: Interview about Experience and Motivation for Rating

All 7 participants agreed they rate because they felt a responsibility to inform others about their experience, particularly with extremely good/bad experiences. One participant said: "*if something is very good, I'd like others to enjoy it too, and if it's bad, I write [a review] so that others won't have such a bad experience.*" In addition, one participant mentioned the difficulty of rating mediocre experiences with professors on ratemyprofessor.com: *"I usually rate the ones that I like the most, and the ones that I don't like at all. […] There's quite a big range of the mediocre ones. You don't exactly know how good they are."*

Further, 5/7 participants said that they rate to improve the system's recommendations. However, 1 participant mentioned that her concern for privacy makes her reluctant to provide information to the system. By contrast, 4/7 participants said that they rate because they have a desire to express their opinion. All of the participants had the desire to influence aggregated ratings but most felt that they had minimal influence to do so when there are many raters; therefore only 3/7 participants agreed that this desire was a motivation. Five of the 7 agreed that they rate to keep a record of their experiences for their own future reference, with 4 of those saying they used them for categorization and organization of their records and collections. Finally, rating can also bring pleasure as one user said "*I think the pleasure of expressing one's opinion reinforces that loose social responsibility.*" However, responses on the fun of rating were mixed. From the study on Movielens users it was concluded that "for at least some users, rating is not a means to an end, but an end of its own"; however, based on our interviews, the fun of rating seem to be a result of the pleasure of achieving other goals such as expressing opinion, and organizing experiences. Figure 2 presents a summary of responses to the 5-point Likert scale questions, binned into three categories: agree, neutral, and disagree.

Part 2: Rating Exercises

Based on our observations of our participants actually rating items, comparison played an important role when rating movies, restaurants, products, and recipes, but not music. When participants were asked to justify their ratings, all of them at some point referred to relevant experiences (e.g., restaurants visited) or similar items (e.g., products owned). Even one participant who seemed to have clear criteria for each star level changed his opinion about the first movie he rated for us after rating a second movie. This implies that those who rate based on specific criteria compare items with respect to those criteria. We expected that those who rate more regularly rely less on direct comparison. However, the interviews showed that despite specific criteria, when rating multiple movies they sometimes went back and adjusted a rating for consistency. In one of the interviews, where the participant rated three movies, she justified her third rating by saying that the movie was between the previous two (rated as 5 and 8) and rated it a 6 since it was closer to the first one.

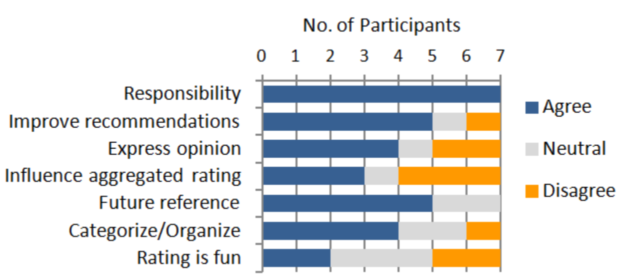

Figure 2. Answers to the Likert scale questions on motivation, binned into 3 categories: Agree, Disagree and Neutral (*N*=7).

We observed a tendency to not use the highest or lowest rating, which is referred to as ends-aversion bias [18]. One user justified her strategy saying that *"I don't have a sense of what [star level] to select, I prefer not to select the highest, because there can always be a better one."*

Another interesting strategy was rating in batches. One participant mentioned that every few weeks, she rates the movies that she watched during that period to have a complete record. Such strategies may shift the trade-off between speed and accuracy of rating.

Summary of Findings

Feeling of responsibility, personal future reference, and improving recommendations were the most common reasons for rating items. While this is fairly consistent with the findings of the Movielens survey [10], we found that the fun of rating is mostly a result of the pleasure of achieving other goals. Overall, the participants did not have absolute and persistent understanding of what is meant by each star level, which is consistent with the literature [3, 7]. All participants recalled related items to decide ratings, even when they had specific criteria in mind for each star level. Moreover, the participants differed from one another in their interpretations of the different star levels. They used different rating strategies and behaviors such as rating in batches, rating only distinct experiences (e.g., extremely good/bad), and not using the lowest and highest ratings (i.e., ends-aversion bias [18]). These differences can add noise to online recommendations and aggregate ratings that commonly inform decisions.

ANALYSIS OF THE DESIGN SPACE

Based on the results from the formative study and the literature on attitude measurement, we identify two axes of interest in the design space (Table 1). The Recall Support axis relates to the use of additional information, such as previously judged items, when making a judgment about a new item. According to the formative study and the literature, people try to recall relevant items. Recalling this information from memory without support may increase cognitive load. We conjectured that an interface that explicitly shows information about previously judged items might facilitate the judgment of new items. In the three-phase cognitive model for reporting attitude [14], this consideration falls into the deliberation phase, where

Recall Support: Items shown	**Measurement Scale**	
	Absolute Rating	Relative Ranking
0	Stars	(Impractical)
log *k*, in pairs		Binary
n	Stars+History	
k		List

Table 1. Design space taxonomy: *n* is the number of levels in the rating system, *k* is the number of items rated or ranked.

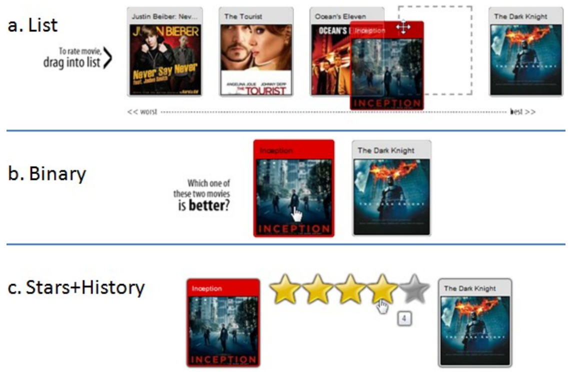

Figure 3. Low-fidelity HTML prototypes: The user is rating the movie Inception. In part (c) the “Dark Knight” thumbnail is displayed when the user hovers over the fourth star.

relevant pieces of information are recalled to form an opinion. Two variables are of interest for this axis: n, the number of levels in the rating system, and k, the number of items that have already been judged. Also of interest are the methods that can be used for providing such information.

The Measurement Scale axis addresses whether the judgment made is an absolute rating or a relative ranking. Our observations and the literature [2,13] suggest that interfaces based on comparison or ranking can be viable alternatives to rating interfaces. The result of ranking all items against each other is a fully ordered list, whereas the result of rating is a partially ordered set with an equivalence class for items at each rating level. Rating is the more familiar and less cognitively demanding form of judgment. Ranking, on the other hand, is intriguing as a solution to the ends-aversion bias [18]. A ranking interface necessarily provides some recall support, because the item to judge is compared against one or more others (it would be practically impossible without this support). Therefore, these two axes are interdependent. In particular, the upper right quadrant (Table 1) is necessarily empty. Our goal in creating an interface in the lower left quadrant, combining rating with recall support, was to untangle the conjectured benefits of recall support from the entailments of fully ordering items via ranking vs. partially ordering them via rating. We also chose to investigate four points along the spectrum of Recall Support possibilities in order to understand whether more is always better.

Depending on the item being judged, different measurement scales may better fit the cognitive process of judgment. For example, in the formative study judging music appeared to be different than other items, in that the participants did not find direct comparison of songs as useful. The Measurement Scale axis is closely related to the response phase of the discussed cognitive model, in which an opinion is transformed to match a response scale.

We generated four design concepts based on these axes. Two used absolute rating, and two used relative ranking. Stars is a standard rating interface, where a judgment for each item is made in isolation with no explicit recall support. Stars+History (S+H) is the same interface with the addition of recall support by showing one example item for each star level. The intent is that the item acts as a reminder for the meaning of that level, serving as a fast calibration mechanism. Binary is a ranking interface, where the user makes a succession of pairwise judgments between two movies, for a total of log k comparisons for each item that is judged (as in binary search). List is also a ranking interface, where the user inserts a movie into the desired location in the list of the k items judged so far. With the two ranking interfaces, comparison is explicitly part of the process. With the S+H rating interface, comparison is available if desired. In all cases, the intention of including comparison is to reduce cognitive load and support accuracy, by reminding the user of previous decisions.

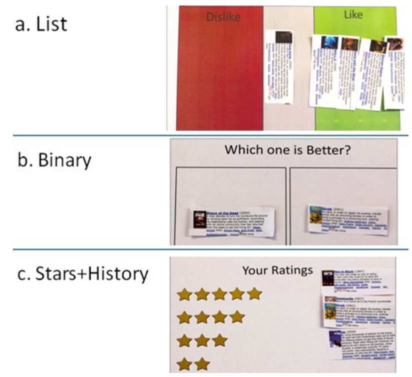

Figure 4. Low-fidelity paper prototypes.

DESIGN PROCESS

After the formative study, the development of the design space taxonomy, we continued with three phases of prototyping. We built low-fidelity HTML and paper prototypes for informal evaluation of the interactive and cognitive aspects of the designs with 10 participants (no overlap with those in formative study). We capped off our process with a medium-fidelity prototype that was used in a formal evaluation with 16 new participants.

We chose to use movies as the subject of judgments, mainly because many people watch movies frequently. Additionally, experience of watching a movie is often measured with a single judgment without breaking it down into multiple components, whereas other items can be judged based on their various components. For example, many products can be judged with respect to their value for money, features, and durability, while restaurants can be judged based on the ambience, service, and food quality.

Low-fidelity HTML and Paper Prototypes

We designed interactive HTML prototypes of the List, Binary, and S+H interfaces (Fig. 3). We collected informal feedback on interaction aspects of the designs from our participants. The main barrier to gaining more insight from our participants was that we hardcoded a small set of 5 movies, and participants who had not watched all of them did not have opinions to express. Thus, these prototypes could not fully support the cognitive process of judgment.

To address this limitation and other feedback, we created low-fidelity paper prototypes where participants could choose 20 movies they had seen from a large set of 150 possibilities. Two major changes were made to the designs, as well as several low-level changes involving labels and layouts. The first major change was partitioning the List interface into three sub-lists of like, neutral and dislike categories (Fig. 4, top). This variant addressed two issues. First, the final output of the List interface was not representative of the user's full opinion: by just looking at the list we could only say that the user liked movie A more than movie B, but there was no way of saying if the user likes movie A or not. Secondly, there were individual differences in the desired level of accuracy. In the formative study, some of the users showed a preference for like/dislike buttons over interfaces that allow for more precision. In the new variant of the List interface, users could just leave the items in the three areas, or rank them within each area. We decided to assess informally both the standard List and the new variant to ensure that the new variant is at least as effective. The second major change following the HTML prototypes was to the presentation of history in the S+H interface. In the initial design, the last movie for each star level appeared only when the user hovered over the corresponding star. Some of the users had to hover over the stars several times before making a decision. We decided to have the history always visible to reduce the required effort.

We asked participants to find movies that they had watched from printouts of popular movies. We then used movies from the same genres for the evaluation of each of the paper prototypes to collect feedback on the judging process. Based on the participants' comments, we decided that the design alternatives were sufficiently refined to allow for a formal evaluation. From the feedback, we came up with more complex design ideas, mostly hybrid designs. However, we decided to keep the prototypes as simple as possible to be better able to relate the results of the final study to the conceptual differences of the prototypes.

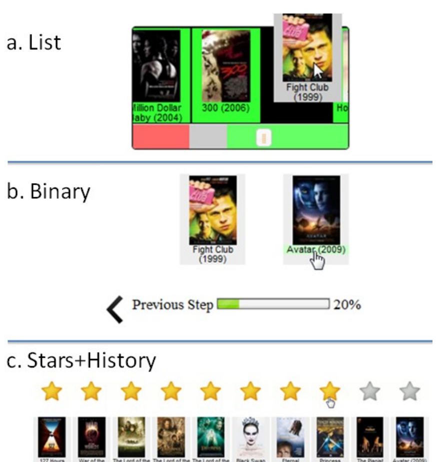

Figure 5. Medium-fidelity HTML prototypes (All thumbnails were identical in size in the prototypes, scaled here only.)

Medium-fidelity HTML Prototypes

We designed medium fidelity prototypes of the three design alternatives. We also built Stars, a standard 10-star interface similar to IMDB.com, giving us four interfaces in total. Although the 5-star interface is widely used for expressing opinions of movies, many of the sites (such as rottentomatoes.com) allow half star ratings, essentially yielding a 10-point scale. The design of medium fidelity prototypes of the Binary interface (Fig. 5.b), and the S+H interface (Fig. 5.c) were basically the same as the corresponding low fidelity paper prototypes. A minor change to the Binary interface was the addition of a progress-bar and navigation buttons for navigating between the comparisons to enable recovery from erroneous clicks.

The List interface (Fig. 5.a) underwent some changes. We added a navigation scrollbar in the form of a scented widget [24], such that three colors showed the distribution of movies into the three categories of like/neutral/dislike. These colors also showed the location of the boundaries between the categories, facilitating navigation of the list. For example, for ranking a barely liked movie, the user can jump to the beginning of the like section then drag and drop the new movie into that part of the list. For the study, we wanted to test the interfaces in a form that would be consistently usable even with a large set of movies, so we chose not to enable auto-scroll while dragging because that technique is suitable only for navigating short distances. In the same spirit, we decided to have a small window into the list (i.e., only show a few movies for context), as showing a large portion of the list of movies at once is not possible with a long list. There was no compelling need to further consider scalability issues for the Binary interface because the number of movies that must be compared to place the movie in a ranked list only grows as the logarithm of the number of items in the list.

FINAL PROTOTYPE EVALUATION

The goal of this experiment was to understand how the conceptual differences in the interfaces affect users' opinion and behavior in using the final medium-fidelity prototypes. We collected both quantitative and qualitative data through an open-ended questionnaire and usage logging, and triangulated the results whenever possible.

Note that in this study the quantitative data were less reliable and informative than the qualitative data; because the ranking interfaces were novel and dramatically different from the standard rating interfaces, they would have needed further development to become sufficiently mature for a fair quantitative comparison. Thus the quantitative analyses need to be interpreted in the context of the qualitative data.

Participants

16 new volunteers (5 females) with various levels of prior experience rating movies participated in the study. They held a variety of occupations including bartender, clerk, secretary, salesperson, engineer, software developer, and students at undergraduate, masters, and PhD levels.

Methodology

Based on the formative study, we knew that a shortcoming of the standard Star rating interface is that users do not have an absolute and persistent understanding of what each star level means; therefore they cannot maintain consistency when judging movies at different times. Thus, in order to enhance ecological validity, we conducted an experiment with four sessions, separated by day-long time intervals. People typically judge items at their leisure; therefore the prototypes were deployed over the Internet, allowing maximum flexibility. A within-subject design was used for this experiment, with interface as the within-subject factor.

Qualitative and quantitative methods were used to offset the weaknesses of one method with the strengths of the other. For example, in this experiment not every participant commented on every issue; however, the extent to which each issue generalizes can be estimated roughly based on the quantitative data. On the other hand, the qualitative data exposed both conceptual and implementation issues that are essential for interpreting the quantitative data.

Task and Procedure

We began by collecting a list of 20 movies from each participant, ones they had watched in the last 3 years. In each session of the experiment, each participant performed a randomized sequence of 20 judgment tasks, where a task consisted of judging a movie using one interface. Each sequence consisted of judging all 20 movies divided into 5 blocks of 4 judgment tasks, one with each of the four different interfaces. By the end of the fourth session, all of the 20 movies had been judged using each of the interfaces. This allowed us to ask participants to compare their performance using each of the interfaces.

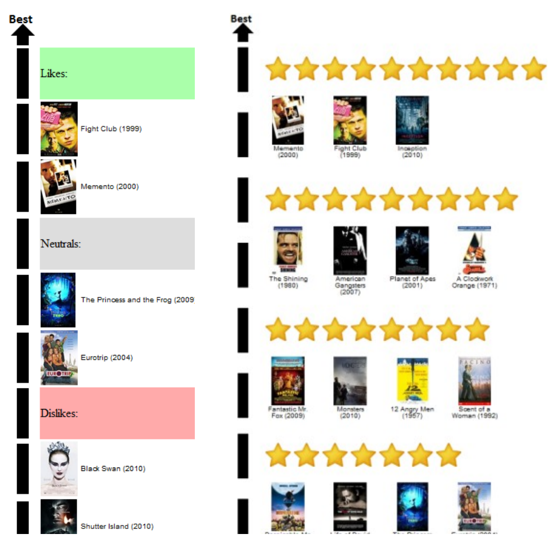

Figure 6. Rating summaries presented to the participants to evaluate the accuracy of the List interface (left) and Stars/ Stars+History interfaces (right). The summary for the Binary interface was similar to the one for the List interface, but without the labels for separating Likes, Neutrals, and Dislikes.

To reduce the effect of remembering judgments from previous blocks, we used the *n*-back distracter task that is commonly used for placing continuous demand on working memory (e.g. [5]). In the *n*-back task "subjects are asked to monitor a series of stimuli and indicate when the currently presented stimulus is the same as the one presented *n* trials previously" [16]. We used *n*=2 with movie pictures as stimuli. We altered the randomization of interface order to ensure that the S+H interface always appeared after the Stars interface in the same block, to avoid having the previous judgments shown by the S+H interface taint users' responses using the Stars interface.

The first session was a practice session in which the interfaces were explained. For each session we sent links to the trials to each participant, and ensured that at least 12 hours had passed from their previous sessions (actual average difference between sessions was 29 hrs). After the last session, we administered an open-ended questionnaire via in-person interviews for the available participants and over the phone or self-administered for the others. The interviews were recorded through note taking and when possible on video. Participants ranked and commented on the interfaces based on the measures below.

Measures

Speed: In order to control for the time spent by participants remembering a movie, we used a two-stage interface. It presented the name and a picture of the movie to be judged; it was only after the user clicked on a button that one of the interface alternatives was revealed. The system recorded the speed by logging the mouse click events, the last of which was assumed to be the end of the interaction. In addition, we asked participants to rank the interfaces based on their perceived speed with each interface.

Accuracy: To evaluate the accuracy of the expressed opinions, we showed users a summary of their judgments from each interface and asked them to identify the changes needed to improve the accuracy of the summaries (samples given in Fig. 6). Because of the substantial differences between the output of the ranking and rating interfaces, it was not meaningful to compare the differences between the summaries. Therefore, we asked participants to rank the summaries based on how well each of them represented their movie taste. Additionally, we asked participants to rate themselves in terms of caring about accuracy and speed of judging movies using a 5-point Likert scale.

Mental Demand: Because the study was conducted through the Internet we did not use methods for measuring cognitive load that assume a lab environment. We asked participants to provide qualitative feedback and rank the prototypes based on the mental demand required to express an opinion.

Suitability for Organization: We asked participants to rank the interfaces based on how suitable they are for keeping track of experiences for future reference or for recommending to others.

Fun to Use: According to the formative study, fun of rating is mostly related to the fun of achieving other goals. However, during the low-fidelity prototype testing we noticed that major differences between interfaces can influence how fun they are to use. Therefore, we asked the participants to rank the interfaces based on this parameter.

Overall Preference: Ultimately, we were interested in knowing which interfaces are preferred and we asked the participants to rank them based on their overall preference.

RESULTS

Quantitative Analyses

The ranking data was analyzed using the Friedman test, and Kendall's coefficient of concordance (W) was used for measuring the agreement between participants. The coefficient (W) ranges from 0 to 1, with higher values indicating a stronger agreement. The p values for the pairwise comparisons are Bonferroni-corrected. The summaries of the rankings are shown in Figures 7 and 8.

The subjective and objective measurements of speed matched well (Fig. 7). The interfaces significantly affected perceived speed ($p<0.01$, $\chi^2_3=19.54$, $W=0.41$), with pair-wise comparisons showing that Stars and S+H were perceived to be significantly faster to use than the Binary and the List interfaces. An ANOVA, with Greenhouse-Geisser correction, comparing actual logged speeds revealed the same effect of interface ($F_{1.98,\ 29.74}=8.12$, $p<0.001$, $\eta_p^2=0.35$) as well as for the pair-wise t-tests.

There were also significant effects of interface on accuracy ($p<0.01$, $\chi^2_3=13.01$, $W=0.27$) and suitability for organization ($p<0.01$, $\chi^2_3=12.69$, $W=0.26$), with pair-wise tests showing that S+H was perceived to have significantly more accurate results and to be more suitable for organization compared to the Binary interface ($p<0.01$). In terms of the relative importance of accuracy, there was a significant preference for accuracy over speed ($\chi^2_2=6.1$, $p<0.05$): 9 participants considered accuracy to be more important than speed of expressing opinion, whereas only one participant believed speed to be more important.

Mental demand ($\chi^2_3=4.50$, $W=0.094$) and fun to use ($\chi^2_3=2.24$, $W=0.047$) were not significantly affected.

There was a significant difference in terms of participants' overall preference ($p<0.01$, $\chi^2_3=12.63$, $W=0.26$). Pairwise tests showed that S+H was preferred over all others ($p<0.05$ for all three).

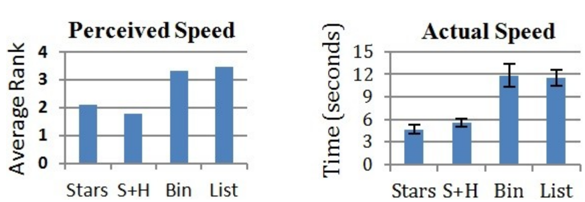

Figure 7. Average ranking of interfaces based on perceived and actual speeds (*N*=16). Note that the y-axes differ. Shorter bars represent faster speeds. Error bars represent Std. Err.

Qualitative Analysis

The primary goal of the qualitative analysis was to examine the hows and whys of the rankings. Therefore, the responses to the open ended questions were reviewed and categorized according to the quantitative measures. Finally, descriptive, meaningful phrases and comments were extracted for reflection, discussion, and integration with quantitative data.

Speed: Several participants made comments regarding the importance of speed, suggesting that the difference in speed of rating between the interfaces was not particularly important. P10 mentioned that *"when I spend 2 hours watching a movie, I don't care about 30 seconds more"*. And P7 said *"I don't care as long as it's reasonable enough… only Binary at the end got so tiring."* P13 went as far as to say: *"Most of the time you're thinking. The 'clicking' doesn't take that much time."*

Several participants explained their speed with the Stars and the S+H interfaces was because of their ease of use. Some of the 7 participants who thought the S+H was faster to use than the Stars interface, talked about how it helped them remember their previous ratings for calibration as P14 said *"everything is in front of you… if you want to rate something similar to this, you just click on it"*. The only participant that felt Stars was faster to use in comparison with S+H said that *"[With Stars] you don't compare, you just say something"* (P13).

Accuracy: Some users felt that the extra information given in the S+H interface affected the level of accuracy they felt they should achieve. P14 said *"If I were given this much information when I'm rating something... I feel like I have to care more about my accuracy, and even with this one [List interface] too … [With Stars interface] I'm not gonna spend time thinking about how accurate am I being. I just rate it; gut feeling…"*, and P7 said *"now that I see the history I do care… it allowed me to care about my accuracy*". Nine participants ranked the S+H interface, the highest. The poor ranking of accuracy of the List interface was due to two factors. First, although it did allow users to rank movies accurately, it also allowed them to simply categorize them into Like, Neutral, and Dislike. Three participants used it only for categorization; therefore their final ranking was a poor representation of their exact taste. The second factor was the interaction effort required to put a movie in the desired spot. Three other participants expected the prototype to support auto-scrolling when

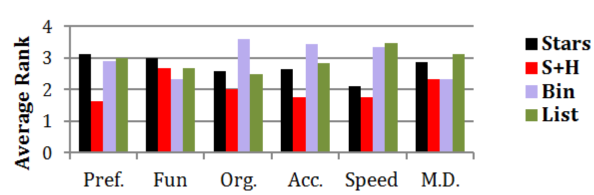

Figure 8. Average ranking of interfaces based on preference, fun, suitability for organization, accuracy, speed, and mental demand (*N*=16). Shorter bars indicate better ranks.

dragging a movie. As mentioned earlier, having auto-scroll would have sacrificed generalisability to interaction with long lists. The participants were explicitly instructed to first use the scrollbar to navigate to the position in the list that they wanted to place the movie, then drag and drop the movie. However, the 3 participants mentioned that they forgot the instructions, which made it hard to place a movie in the right position, as it required multiple drag and drops. Consequently, sometimes they sacrificed accuracy. Regarding the poor ranking of accuracy of the Binary interface, several participants indicated that when clicking quickly they might have clicked incorrectly, and sometimes when the two movies were not easily comparable, their decisions might not have been accurate.

Mental Demand: One of the 4 participants who preferred both ranking interfaces over rating interfaces said *"you don't have to quantify anything, you just sort them."*(P13) Several participants indicated that they try to be consistent with their ratings and to remember their previous ratings, and that S+H facilitated the process. P4 compared the Stars and S+H interface saying that *"with Stars you have to think about what you rated the previous ones. What's the definition of a 7 and an 8? [With S+H] you just get reminded what the definition of a 7 and an 8 is."* Seven others shared this preference while 4 ranked them the same, and 4 preferred the Stars interface.

In regards to the underwhelming ranking of the List interface (though not significant), 2 participants mentioned that they were trying to compare the new movie with several movies at a time, whereas with the Binary interface it was easy to compare only one movie. P2 mentioned:*"I really like the simplicity of the Binary, it doesn't require a lot of thinking and the results are calculated for you in the end."* and P13 said" *"[With Binary interface] you don't have to compare one thing to the whole [list] at the same time"* P2, P13 and 4 others found Binary the least mentally demanding. On the other hand, several participants mentioned the difficulty of comparing two movies whose merits they considered incomparable. For example, P10 said *"Sometimes they are not comparable ... One is funny, one has a great story".*

Suitability for Organization: According to the participants, two main factors determine the suitability of interfaces for organization: first, the accuracy of the opinion captured by the interfaces and second, the preference for having a ranked list or having multiple categories. One of the 9 participants who ranked the Binary interface the last (P7) talked about a problem with the list of movies generated using the Binary interface: *"...if you're a frequent movie goer, you gotta forget [the movies you've watched] and there is no line that draws ok up to this point is the ones that I like."* Some of the users appreciated the precision of organization supported by the List interface, while others preferred to have categories, as in star interfaces. For example P7 said: *"The categorization helps me a lot more than the sorting. When somebody says can you recommend me a movie, it's a lot easier to just pull out ... the movies that I've rated 10 of 10"*, and P9 said *"Categories are suitable because for recommending to others it's not important to be accurate."* P7, P9 and 5 others ranked both rating interfaces higher than the ranking interfaces.

Fun to Use: Several participants did not consider any of the interfaces to be fun to use, saying that *"don't know if 'fun' is something you should use to describe a rating system."* (P6) and *"None of these were really fun to use… but at least S+H was visually appealing"* (P12). These participants responded to this question based on the ease of use or low mental demand. Seven users ranked the Binary interface the highest and mentioned that the comparisons in the binary interface were *"almost like playing a game"* (P10) or *"like competition between movies"* (P9).

Overall Preference: Participants talked about the various aspects of interfaces that influenced their preference. The 10 participants who preferred the S+H interface talked about its low cognitive load and ease of use. P7 pointed out the familiarity bias: *"We are so used to stars, so I don't have to think about it"*. The 3 participants who preferred the Binary interface, mentioned its simplicity and fun. The 2 participants who preferred the List interface talked about the fine granularity it supports and their preference for ranking over rating. P1, the only participant who preferred the Stars interface, stated each movie should be *"ranked in its own right, rather than in comparison to others…You rate them in reference to which you would prefer to watch and not exactly comparing them.*"

DISCUSSION AND FUTURE WORK

As discussed earlier, the ranking interfaces need further development to become sufficiently mature for a fair quantitative comparison. We thus weigh our qualitative results more heavily. In this section we discuss those results and propose hypotheses and directions for future research.

Our work shows that augmenting a standard rating interface with previously judged items to improve recall can support expressing opinion without sacrificing speed. The S+H interface provided support for recall while taking advantage of familiarity with the standard Stars interface. It did not require users to express their opinion with the conceptually different method of ranking. The significantly higher preference for the S+H and the qualitative data suggest that providing support for recall can be even more helpful in long term usage. The S+H interface was as accurate or more accurate than the other interfaces, and was as fast as the standard Stars interface and significantly faster than the other interfaces. However, a speed-accuracy tradeoff might come into play for a different set of interfaces; participants reported caring more about their accuracy than their speed.

Contrary to our expectations, we did not observe any significant effect of interface on mental demand. The qualitative data provides some explanation for this. Several participants mentioned that having the previously judged items available made them care more about accuracy. In fact, some of those who did not seem to bother recalling relevant information when using the Stars interface, tried to improve their judgment accuracy when using interfaces that provided recall support. We hypothesize that the increase in effort on accuracy masked any reduction of mental demand.

The underwhelming performance of the Binary and List ranking interfaces is not solely due to the immaturity of the design or familiarity bias in favor of rating interfaces. A fundamental problem with the ranking interfaces was that people had difficulty when asked to compare movies that they considered incomparable. This problem was most severe for the Binary interface which required direct comparisons. Several participants mentioned arbitrary decisions when comparing seemingly incomparable movies and inaccuracy when clicking quickly. Every arbitrary or inaccurate comparison can influence future insertions; when the list goes out of order, new insertions using binary search will be subject to error. Moreover, because the Binary interface shows information about only one pair of movies at a time, users are unlikely to notice an error with a previous comparison. Nevertheless, we believe the ranking interfaces have merits that call for further investigation and design efforts. Specifically, the fun of using the Binary interface and the potential for organizing experiences using the List interface are two avenues for future research.

The Binary and List interfaces generated different results, and interestingly both results often seemed unsatisfying to users. The dissatisfaction might be due to a fundamental mismatch in the mathematical models at play in the Measurement Scales axis: ranking interfaces enforce a fully ordered result set, but people might have a mental model that is only partially ordered. Further studies are needed to investigate this hypothesis, and to explore what mental models people have for other data types beyond the movies that we studied. A related finding is that opinion measurement interfaces are not automatically improved by simply moving higher on the Recall Support axis; List did not dominate S+H or Binary, even though it provided more previously judged items. Further studies could investigate the implications of the relative sizes of the number of rating levels n and the number of items to rate k. Moreover, the method used for selecting which of the previously judged items to be shown may play an important role in the effectiveness of the recall support. The selection of the previously judged items can be done prior to interaction with the user, as in the S+H interface, or based on user input while judging, as in the Binary interface. Alternatively, all judged items can be available, allowing user navigation as in the List interface.

Another limitation of the summary created through the Binary interface is that there are no boundaries that reflect the user's value judgments: there is no information about the boundaries between those disliked or liked by the user. The combination of all of these issues led to the poor ranking of the Binary interface with respect to accuracy and suitability for organization. On the other hand, many of the participants had fun using the Binary interface, which suggests that it might be usable to collect bits of information about people's taste without necessarily using it as the primary interface for recording experiences.

A design challenge for the List interface was scalability. The two strategies of using a small window into the list, and not supporting auto-scroll were devised to ensure that the overall interaction would not be significantly dependent on the number of items judged; however, they slowed down the participants. Moreover, information visualization techniques can be used to facilitate the navigation of long lists [6]. For the Binary interface, scalability is not a major concern because the number of comparisons grows slowly. One possibility for decreasing the number of comparisons is to allow the user to first select the appropriate part of the list, and then use binary search within that area.

Both qualitative results and the poor agreement between participants (based on Kendall's W values) highlight the role of individual differences in opinions about the interfaces, and signify an opportunity for future research on personalization of rating interfaces. One possibility is that there are different levels of willingness by users to take advantage of information about previously judged movies: some ignore it, while others exploit it. Another possibility is that people have varied success in using the information, perhaps, depending on the relevance and accuracy of their previous judgments.

A limitation of all our prototypes is in requiring a name and a visual representation of items. While not every item can be represented visually or with a short representative text, many items such as products or services have icons/images representing them. Thumbnails representing features of the items are widely used. Nevertheless, it may be impossible to create thumbnails for abstract concepts. Showing a small image facilitated recognition, and future studies should assess the applicability of our findings to judging abstract concepts or items that do not have visual representations.

Our taxonomy is but a first step. Several dimensions of the design space yet to be explored include categorical vs. ordered scales, diverging vs. sequential scales, continuous vs. discrete scales, and the amount of precision. The popular approach of tagging can be considered as judgments on a categorical scale, whereas both of the measurement scales we used were ordered. Our scales were also discrete, but precision can be high enough to be perceived as continuous as in computer-based visual analogue scales [20]. Although previous research has shown little difference in reliability and discrimination power for precise 101-point vs. 9-point scales [19], we hypothesize that providing recall support would enable users to benefit from the precision of those scales.

SUMMARY AND CONCLUSION

Our taxonomy and investigation focused on two dimensions of the design space of opinion measurement interfaces: the Measurement Scale and the Recall Support. In our mixed-methods study, users preferred the S+H interface that provided recall support with examples for each rating level, but did not require direct comparisons. Although people compare movies to judge them, this process turned out to be complex. It involves comparison with movies that are related based on criteria largely determined by the movie, and the judge's own viewpoint and experiences.

We identified a number of directions for future research, including the investigation of mental models used when judging various types of items, personalized opinion measurement interfaces, recall support for judging abstract items, recall support for long rating scales, and the effect of recall support on effort for accuracy. We document the need for further development of the ranking interfaces, including design advances that will leverage the specific merits of the ranking interfaces (e.g., fun of using the Binary interface).

In addition to the pervasive use of opinion measurement interfaces for expressing opinions about products and services, researchers in various disciplines also use Likert scales and other simple interfaces to elicit people's opinion. Studying and advancing interfaces for measuring opinions should result in more accurate and internally valid subjective data, thereby improving the results of those research projects that rely on them. There is a lot more to be learned about various aspects of opinion expression and representation interfaces. We believe that other dimensions of the design space as well as the design concepts presented here deserve further investigation and the goal of this paper was to stimulate discussion on this topic, not to conclude it.

REFERENCES

1. Abeyratna, S., Paramei, G., Tawfik, H., and Huang, R. An affective interface for conveying user feedback. *Proc. Computer Modeling and Simulation*, IEEE Computer Society (2010), 369-374.
2. Alwin, D.F. and Krosnick, J.A. The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly 49*, 4 (1985), 535 -552.
3. Amatriain, X., Pujol, J., and Oliver, N. I like it... I like it not: Evaluating user ratings noise in recommender systems. *Proc. UMAP 2009*. Springer (2009), 247-258.
4. Amatriain, X., Pujol, J.M., Tintarev, N., and Oliver, N. Rate it again: increasing recommendation accuracy by user re-rating. *Proc. RecSys 2009*, ACM (2009), 173-180.
5. Cades, D.M., Trafton, J.G., Boehm Davis, D.A., and Monk, C.A. Does the difficulty of an interruption affect our ability to resume? *Proc. Human Factors and Ergonomics Society Annual Meeting 51*, (2007), 234-238.
6. Cockburn, A., Karlson, A., and Bederson, B. A review of overview+detail, zooming, and focus+context interfaces. *ACM Computing Surveys 41*, (2009), Article 2.
7. Cosley, D., Lam, S., Albert, I., Konstan, J., and Riedl, J. Is seeing believing?: How recommender system interfaces affect users' opinions. *Proc. CHI 2003*, ACM (2003), 585-592.
8. Eagly, A.H. and Chaiken, S. *The Psychology of Attitudes.* Wadsworth, Belmont, CA, USA, 1993.
9. Gawronski, B. Attitudes can be measured! But what is an attitude? *Social Cognition 25*, 5 (2007), 573-581.
10. Harper, F., Li, X., Chen, Y., and Konstan, J. An economic model of user rating in an online recommender system. *Proc. UM 2005*. Springer (2005), 307-316.
11. Herlocker, J., Konstan, J., Borchers, A., and Riedl, J. An algorithmic framework for performing collaborative filtering. *Proc. SIGIR 1999*, ACM (1999), 230–237.
12. Ivanov, A., Erickson, T., and Cyr, D. Plot-polling: Collaborative knowledge visualization for online discussions. *Proc. Intl. Conf. Information Visualization 2006*, IEEE Computer Society (2006), 205-210.
13. Klein, M., Dülmer, H., Ohr, D., Quandt, M., and Rosar, U. Response sets in the measurement of values: A comparison of rating and ranking procedures. *Intl. J. of Public Opinion Research 16*, 4 (2004), 474 -483.
14. Krosnick, J.A., Judd, C.M., and Wittenbrink, B. The Measurement of Attitudes. In *The Handbook of Attitudes*. 2005, 21–76.
15. Lietz, P. Research into questionnaire design. *Intl. J. of Market Research 52*, 2 (2010), 249–272.
16. Owen, A.M., McMillan, K.M., Laird, A.R., and Bullmore, E. N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping 25*, 1 (2005), 46-59.
17. Ozakca, M. and Lim, Y.-K. A study of reviews and ratings on the Internet. *Ext. abstracts CHI 2006*, ACM (2006), 1181–1186.
18. Philip, B.R. Generalization and central tendency in the discrimination of a series of stimuli. *Canadian Journal of Psychology 1*, 4 (1947), 196-204.
19. Preston, C.C. and Colman, A.M. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica 104*, 1 (2000), 1-15.
20. Reips, U.D. and Funke, F. Interval-level measurement with visual analogue scales in Internet-based research: VAS Generator. *Behavior Research Methods 40*, (2008), 699-704.
21. Van Schaik, P. and Ling, J. Design parameters of rating scales for web sites. *ACM Transactions on Computer-Human Interaction 14*, 1 (2007), Article 4.
22. Thurstone, L.L. Attitudes Can Be Measured. *The American Journal of Sociology 33*, 4 (1928), 529-554.
23. Turnbull, D. Rating, voting & ranking: designing for collaboration & consensus. *Ext. abstracts CHI 2007*, ACM (2007), 2705–2710.
24. Willett, W., Heer, J., and Agrawala, M. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (2007), 1129-1136.
25. Ye, Q., Law, R., and Gu, B. The impact of online user reviews on hotel room sales. *Intl. J. of Hospitality Management 28*, 1 (2009), 180-182.