Exploring the Design Space of Rating Interfaces

First Author Name (Blank if Blind Review) **Second Author Name (Blank if Blind Review)**

ABSTRACT

Star rating interfaces are widely used on the Internet for quantifying people's feelings towards a product or service. However, there is little known about the effectiveness of these interfaces and the design space of rating interfaces is relatively unexplored. In this paper we present an exploration of the design space through a formative study of how and why people rate items, followed by iterative prototyping of three alternative interfaces for eliciting quantitative opinions. Based on the formative study, we focused on providing context through comparison in rating and ranking interfaces. We present the results of a multi-session study of the design alternatives with respect to accuracy, speed, fun, mental demand, and users' preference. Providing relevant contextual information, particularly in the form of history of ratings, was appreciated by users and helped them rate more accurately without sacrificing speed. Users preferred to maintain the simplicity of rating interfaces; however our qualitative results showed that individual differences play a prominent role in preference and usage of the interfaces.

ACM Classification: H5.2 [Information interfaces and presentation]: User Interfaces. - Graphical user interfaces.

General terms: Design, Human Factors

Keywords: rating, ranking, opinion, attitude, measurement

INTRODUCTION

Rating products and services on the Internet is pervasive and star rating interfaces are the most commonly used mechanism (Figure 1); despite this, little HCI research has investigated interfaces to support rating. Work to date has mostly focused on the parameters and visual design of n-point Likert scale interfaces (e.g. [1,20,22]), or tailored application-specific designs that, while being more expressive, require significant investment for reuse in other application domains (e.g. [13]). Customer attitudes towards products and services play an important role in communication 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 [7,23].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI 2012, May 5-10, 2012, Austin, TX, USA.

Copyright 2012 ACM xxx-x-xxxx-xxxx-x/xx/xx...$10.00.

Figure 1. Rating interfaces in common use today.

Studies of rating interfaces have been very limited in the HCI literature; however, social psychologists and marketing researchers have long been interested in analyzing various aspects of methods for capturing and understanding people's attitudes [15]. We conducted a formative study to expand our understanding of how people use current rating interfaces and why they rate products and services on the Internet. Based on findings from the literature in social psychology and HCI as well as our formative study, we iteratively designed three alternative interfaces that covered both rating (assigning absolute values) and ranking (relative evaluation by placing items in order). Finally, we conducted a mixed-methods study collecting both qualitative and quantitative data about people's opinions and performance with each of the interfaces.

The main contribution of this work is the iterative design and evaluation of several design alternatives for eliciting quantitative opinions. To our knowledge, this is the first reported exploration of the design space of interfaces used to elicit quantitative subjective opinions. Our study showed that users do take advantage of the additional contextual information when provided by rating interfaces to express their opinion more accurately. We also learned that interfaces should provide an easy way of capturing a user's opinion, and further that users care more about their accuracy than the time it takes to express an opinion. However, the complexity of interaction can prevent users from taking advantage of the extra level of accuracy supported by an interface. Ultimately, different rating interfaces are appropriate for different purposes and different users. In the following section we briefly explore the literature on attitude measurement and the current practices.

ATTITUDE MEASUREMENT

Although there is no agreement on what an attitude is [9,15], one of the often-cited definitions given by Eagly and Chaiken [8] is: "tendencies to evaluate an entity with some degree of favor or disfavor, ordinarily expressed in cognitive, affective, and behavioral responses". This definition is essentially based on Thurstone's view of the concept in his seminal work on measuring attitudes [21]. Thurstone differentiated attitude from opinion, by defining opinion as an expression of attitude and McNemar describes different views on this distinction [16]. Although this distinction would be desirable for discussing certain aspects of the two concepts, we found it unnecessary for the purpose of this study and the two words are used interchangeably in the rest of this paper.

One approach to designing interfaces for expressing opinion is through understanding the cognitive process that is used to generate opinions. One of the cognitive models of reporting attitudes describes it as a three stage process [15]. 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, in which 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. Based on this model, opinion expression interfaces should help users by supporting the second phase by providing relevant contextual information, and the third phase by asking for external representations that match with the output of the previous phases of the process. Pioneers of attitude measurement believed that attitude cannot be described by any single numerical index and have proposed using several questions to capture attitude. For example, the Likert scale originally referred to the sum of responses on several Likert items [15]. However, currently the prevalent way of measuring attitude is to use single questions, likely for simplicity.

Current online reviewing systems use *n*-point scales to capture and present users' attitudes and $n > 2$ is the most common practice. For example, systems such as iTunes and Amazon use 5-point ratings. Often n is an odd number, so that the mid-point represents neutral and higher and lower positions represent varying positive and negative attitudes. There are also systems that only allow for positive attitudes such as Michelin guide's 3-point and Facebook's 1-point scale (Like). Another common practice is using 2-point scales such as on YouTube (thumbs up/down). An absolute assessment scheme is what these systems have in common, which allows for a quick and easy way of expressing opinions, and displaying aggregated results.

The relative merits of ranking (relative evaluation by placing items in order) and rating (assigning absolute values) mechanisms for measuring people's attitudes has long been a subject of debate [2,14] 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 assessment is choosing an option, ranking can be preferable as it matches well with the concept of choice; however, if the goal is to categorize a set of items, rating the items can be more appropriate.

Rankings are often more cognitively demanding and require concentration, which is problematic when dealing with a long list of items. The prevalence of using ratings instead of rankings has been mainly to reduce phone survey completion time; however, making the task easier may reduce the precision of distinctions between items [2]. Moreover, the lack of consistency in ratings 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.

DESIGN PROCESS

Our design process consisted of four phases. We started with a small formative study to understand how and why people use rating interfaces, and continued with three phases of prototyping, starting with low-fidelity HTML and paper prototypes for informal evaluation of the interactive and cognitive aspects of the design concepts. We capped off our process with a medium-fidelity HTML prototype that was used in a final formal evaluation.

Formative Study

To our knowledge, despite the literature mentioned above on attitude measurement, there is little empirical evidence as to why and how people rate products and services. Ozaka and Lim [18] showed that people tend to give feedback when they have strong opinions. Harper et al. [10] conducted a survey on the Movielens recommender system, to understand users' motivations for rating movies. They found that improving recommendations, fun of rating, and keeping a list of watched-movies are the most important motivators. That study focused on frequent movie raters only. By comparison, we conducted our interviews with a more diverse population in terms of experience with rating systems, and we collected more qualitative data.

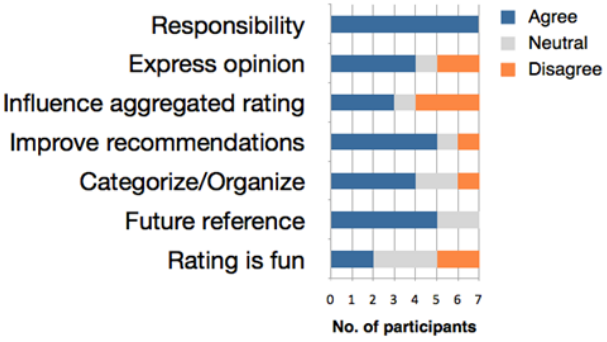

Figure 2. Answers to the Likert scale questions on motivation, binned into 3 categories: Agree, Disagree and Neutral (*N*=7).

We performed our formative study through interviews with 7 subjects (4 females) with various levels of rating experience. The interviews took 25-45 minutes to complete. In the first part, we interviewed participants about their previous experience with rating, with a focus on their motivations. In the second part, we asked them to think-aloud to expose the mental process involved in rating items.

Part 1: Motivation for Rating

We interviewed participants about their previous experience with rating 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. Figure 2 presents a summary of responses to the 5-point Likert scale questions that are binned into the three categories of agree, neutral, and disagree.

All of the 7 participants agreed they rate because they felt a responsibility to inform others about their experience, particularly with extremely good/bad experiences. Regarding rating products, 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 to a sense of responsibility as a reason for sharing extremely good or bad experiences, 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 rated to improve the system's recommendations for them. However, 1 participant mentioned 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. When asked whether the desire to influence aggregated ratings was a motivation, all of the participants felt that this desire existed, but that it is hard to influence when the number of raters is high; therefore 3 of them disagreed that this was a motivation. On a different point, 5/7 agreed that they rated to keep a list of their experiences for their own future reference, and 4 of those also said 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, the responses on the fun of rating were mixed. Based on 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.

Part 2: Rating exercises

In the second part of the formative 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 (IMDB, Yelp, YouTube, etc.). We used a think-aloud protocol, and probed if further explanation of the reasoning behind their rating was needed. Based on our observations, comparison played an important role when rating movies, restaurants, products, recipes but not music. When participants were asked to justify their ratings, all of them at some point referred to other experiences or similar items. Even a participant who seemed to have clear criteria for his ratings 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 more specific criteria, when rating multiple movies they often went back and adjusted a rating to be consistent. 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. Overall, the participants did not have absolute and persistent understanding of what is meant by each star level.

An interesting strategy we observed was users not giving the highest or lowest rating. One of the users 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."* This heuristic makes the extreme ends of the scale useless for some users, and thus converts a 5-point scale to a 3-point scale. On a different note, one participant commented on her rating habit that every few weeks, she rates the movies that she had watched during that period to have a complete record.

Summary of Findings

Feeling of responsibility, personal future reference, and improving recommendations were the most common reasons for rating products and services. All of the participants recalled related items or experiences to decide ratings, even if they had criteria for each star level. The participants had different interpretations of star levels and used different rating strategies such as rating only distinct experiences (e.g. extremely good/bad), not using the lowest and highest ratings, and rating in batches. Some of these differences can introduce noise to automatic recommendations and aggregate ratings that are commonly used in a decision making process.

Design Alternatives

We chose to use movies as the subject of our rating interfaces, mainly because many people watch movies frequently. Additionally, experience of watching a movie is often rated without breaking it down into multiple aspects, whereas other products or services can be rated based on their various aspects. For example, many products can be rated with respect to their value for the money, features, and durability, while restaurants can be rated with respect to the ambience, price, service, and quality of food.

Based on the motivations seen for using rating interfaces and the literature on alternative methods of eliciting users' opinion, we sketched several design ideas and collected preliminary informal opinions on them from potential users. The main criterion for selecting the design sketches was the simplicity of the mental model suggested by the interface, ease of interaction, and accuracy of capturing opinion. We wanted to cover both rating and ranking in our design options. Ranking would not be plagued with the problem of people not using highest/lowest values, while rating is more familiar to people. To support consistency in rating we also wanted to use designs that leveraged comparison, so as to give context and help anchor the user's ratings. As discussed earlier, based on the cognitive model for reporting attitude, providing context can facilitate the deliberation phase, and ranking mechanisms affect both the deliberation and the response generation phases [15]. Informed by the theory and the findings of the formative study, the three design ideas that we next investigated used comparison in some way: Stars+History, Binary, and List (shaded squares in Table 1). The Stars+History and the List interfaces both bring more context to the rating through viewing multiple previously rated items at one time. The Binary interface offers only one comparison at a time, and as such gives a focused view.

Low-fidelity HTML and Paper Prototypes

We designed interactive HTML prototypes of the three most promising design ideas. The List interface (Figure 3.a) presents a ranked list of movies and allows the user to place the new movie into the list. The Binary interface (Figure 3.b) enables creating a ranking through a series of binary comparisons, in which the user selects one of the two presented movies at each comparison. We considered several possibilities for selecting the movies to be compared with the new movie and, ultimately, we decided to use the binary search algorithm to find the right place for the movie in the ranked list. The Stars+History interface augments standard n-star interfaces by showing the last rated item for each star level, when hovered over. (Fig 3.c shows that the "Dark Night" thumbnail is displayed when the user hovers above the fourth star.)

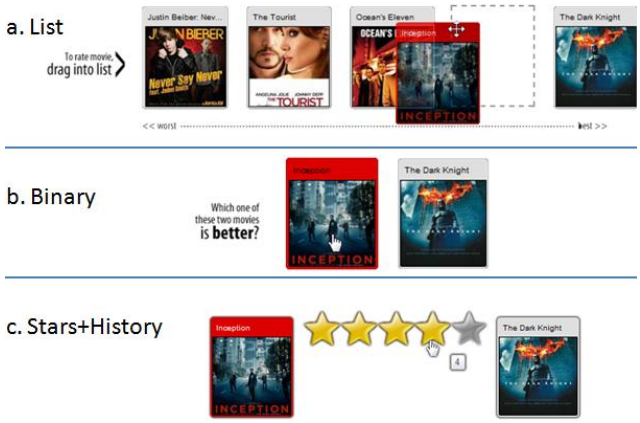

Figure 3. Low-fidelity HTML prototypes: The user is rating

Table 2. The design space considered: The shaded area represents the interfaces that leverage comparison.

	Absolute (rating)	Relative (ranking)
Focus only	Stars only	Binary
Focus+Context	Stars + History	List

We collected informal feedback on interaction aspects of the designs. However, the HTML prototypes could not fully support the cognitive process of evaluation and opinion expression, because the participants had not watched some of the movies coded in the prototypes. To address this limitation and other feedback, we created low-fidelity paper prototypes. Two major changes, as well as several low-level changes (e.g. labels and layouts), were made to the designs. The first major change was partitioning the List interface into three sub-lists of like, neutral and dislike categories (Figure 4, top). This variant of the List interface addressed two issues. First, the final output of the List interface was not representative of the user's 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 are individual differences in the desired level of accuracy. Some of the users preferred to use like/dislike buttons instead of interfaces that allow for more accuracy. 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 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 Stars+History interface. In the initial design, the last movie for each star level appears only when the user hovers 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 for seeing it.

In order to investigate the cognitive process of evaluation and opinion expression we asked participants to find movies that they had watched from printouts of 150 popular

Figure 4. Low-fidelity paper prototypes.

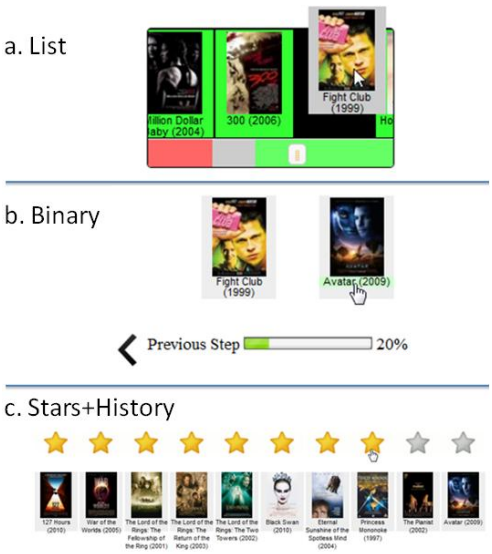

Figure 5. Medium-fidelity HTML prototypes.

movies from three genres. We then used movies from the same genres for the evaluation of each of the paper prototypes to collect feedback on how the users were making rating decisions. Based on the participants' comments, we decided that the design alternatives were sufficiently refined to allow for a more formal exploration of the design space. Based on the feedback from both the HTML and paper prototype evaluations 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.

Medium-fidelity HTML Prototypes

We designed medium fidelity prototypes of the three design alternatives and also built a standard 10-star interface (hereafter called Stars interface) similar to IMDB.com[1], thus four designs in total. The design of medium fidelity prototypes of the Binary interface (Figure 5.b), and the Stars+History interface (Figure 5.c) were essentially 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 (Figure 5.a) underwent some changes. It provides a focus+context view of the ranked list of movies and the scrollbar (the 3-color band at the bottom) can be used to navigate 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 the list. In order to ensure that the usability of the List interface will not diminish as the number of movies goes up, auto-scroll while dragging, a feature that is suitable for navigating short distances while dragging, was purposefully not supported. Additionally, we decided to have a small window into the list (i.e., only show a few movies for context), because showing a large portion of the list of movies at once is not possible with a long list. Unfortunately we were not able to devise strategies similar to those described above for the Binary interface which is also subject to scalability challenges; however, in binary search the number of comparisons grows very slowly (it is $log_2(number\ of\ movies)$); consequently, the interaction time required for placing a movie in a ranked list using the Binary interface will increase slowly.

EVALUATION OF MEDIUM-FIDELITY PROTOTYPES

The goal of this experiment was to explore the design space and understand how the conceptual differences in the interfaces affect users' opinion and behavior. We collected both quantitative and qualitative data through interviews and usage logging, and we triangulated the results whenever possible.

Participants

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

Methodology

Based on the formative study, one of the shortcomings of the standard n-star interface is that users do not have an absolute and persistent understanding of what each star level means; therefore they cannot maintain consistency when rating movies at different times. Thus, in order to enhance ecologically validity, we conducted an experiment with four sessions, separated by (approximately) day-long time intervals. Moreover, people typically rate items at their leisure; therefore we allowed participants to work with the interfaces whenever and wherever they wanted. The prototypes were deployed over the Internet, allowing for maximum flexibility. A within-subject design was used for this experiment with interface as the within-subject factor.

Task and Procedure

Before starting the experiment, we collected a list of 20 movies from each participant, ones that they had watched in the last 3 years. During the experiment, each participant performed a randomized sequence of 20 rating tasks in each session, where a task consisted of rating a movie using one interface. Each sequence consisted of rating all 20 movies divided into 5 blocks of 4 rating tasks, one with each of the four different interfaces. Therefore, in every session users used each interface to express their opinion about 5 movies, and by the end of the fourth session, all of the 20 movies had been rated using each of the four interfaces. This allowed us to ask participants to compare their performance using each of the interfaces.

[1] The 5-star interface is also widely used for expressing opinions of movies, but usually allows half star ratings (such as on rottentomatoes.com) essentially making it a 10-point scale.

We used a distracter task that placed demand on working memory between blocks of trials to reduce the effect of seeing other ratings in the previous block of trials. The n-back task 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 to indicate when the currently presented stimulus is the same as the one presented *n* trials previously" [17]. We used the *n*-back task (*n*=2) with movie pictures as stimuli. The Stars+History interface shows the ratings of several movies which could have been potentially problematic if the Stars interface appeared after it and asked about one of those movies. Therefore, we altered the randomization of the order of interfaces in a way that the Stars+History interface never appeared before the Stars interface in the same block.

The first session was considered a practice session, in which the experimenter was available (physically or on the phone) to explain the interfaces. After obtaining consent for their participation, each of the interfaces was explained and all questions were answered. After the first session we sent emails, including a link to the first trial of the session, to each of the participants. Before sending the emails, we ensured that at least twelve hours had passed from when the user had finished the previous set. The average of the time difference between the sessions was 29 hours.

After the last experiment session, we interviewed the participants to collect their opinion on various aspects of the interfaces. Participants were asked to comment on and rank the interfaces based on accuracy, speed, fun, suitability for organizing experiences, and overall preference. We decided to use ranking of interfaces instead of rating, to ensure that the participants would be able to differentiate their opinion about the interfaces; however, we allowed the participants to rank two interfaces with the same rank, to avoid imposing artificial differences. We interviewed the 11 physically available participants and asked the other 5 remote participants to fill out a questionnaire designed based on the interview script.

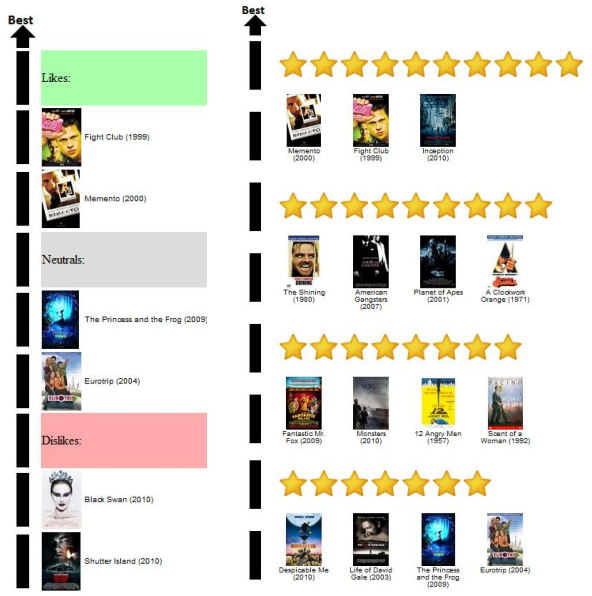

Figure 6. Rating summaries presented to the participants to evaluate the accuracy of the List interface (left) and Stars/ Stars+History interfaces (right). The summary for the Binary interface was similar to the one for the List interface, but without the labels for separating Likes, Neutrals, and Dislikes.

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 rated; it was only after the user clicked on a button that a rating interface 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, as part of the post-test interview, we asked participants to rank the interfaces based on their speed with each interface.

Accuracy: To evaluate the accuracy of the expressed opinions, we showed users a summary of their ratings from each interface and asked them to identify the changes needed to improve the accuracy of the summaries, samples shown in Figure 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 rating movies using a 5-point Likert scale. This was to see if it was possible to explain their overall preference based on their perceived accuracy and speed with the interfaces.

Mental Demand: Due to the nature of the study, we were not able to use lab-specific methods for measuring cognitive load. Therefore, we asked participants to provide qualitative feedback and rank the prototypes based on the mental demand to express an opinion.

Suitability for Organization: According to previous research [10] and our formative study, one of the main motivations for using rating interfaces is to keep track of experiences for future reference or for recommending to others. We asked participants to rank the interfaces based on their suitability for organization.

Fun to Use: According to our preliminary study, fun of rating is mostly related to the fun of achieving other goals such as expressing opinion. However, during the experiments with the low-fi prototypes we noticed that major differences between interfaces can cause different levels of fun of usage. Therefore, we asked the participants to rank the interfaces based on how fun they were to use.

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 (Kendall's W) is used for measuring the agreement between participants. The coefficient (W) ranges from 0 to 1, with higher values indicating a stronger agreement. P values for the pairwise comparisons are Bonferroni-corrected. The summaries of the rankings are shown in Figures 7 and 8.

Speed: The subjective and objective data matched fairly well, as seen in Figure 7. The interfaces did significantly affect perceived speed ($p<0.01$, $\chi_3^2=19.54$, $W=0.41$), with pair-wise comparisons showing that Stars+History was significantly faster to use than the Binary ($p<0.05$) and the List ($p<0.01$) interfaces and the Stars interface significantly faster than the List interface ($p<0.05$). An ANOVA, with Greenhouse-Geisser correction, comparing actual logged speeds revealed the same effect of interface ($F_{1.98,\ 29.74}=8.12$, $p<0.001$, $\eta_p^2=0.35$) and results for pair-wise t-tests .

Accuracy: There was also a significant effect of interface on accuracy ($p<0.01$, $\chi_3^2=13.01$, $W=0.27$), with pair-wise tests showing that Stars+History resulted in significantly more accurate results compared to the Binary interface ($p<0.01$), but no other comparisons were significant. In terms of the relative importance of accuracy, there was a significant preference for accuracy ($\chi_2^2=6.1$, $p<0.05$): 9 participants considered accuracy to be more important than speed of rating, whereas only one participant believed speed to be more important.

Suitability for Organization: Interface also significantly affected the suitability for organization ($p<0.01$, $\chi_3^2=12.69$, $W=0.26$), and pair-wise tests showed that the effect is mainly caused by the significant difference of the Binary interface and the Stars+History interface ($p<0.01$).

Overall Preference: The interfaces did have a significant impact on participants' overall preference for ranking interface ($p<0.01$, $\chi_3^2=12.63$, $W=0.26$). Pairwise tests showed that Stars+History interface was significantly preferred over the Stars ($p<0.01$), Binary ($p<0.05$) and List ($p<0.05$) interfaces. Only one participant ranked the Stars interface as the overall best interface and mentioned that it was because of familiarity and simplicity of the Stars interface.

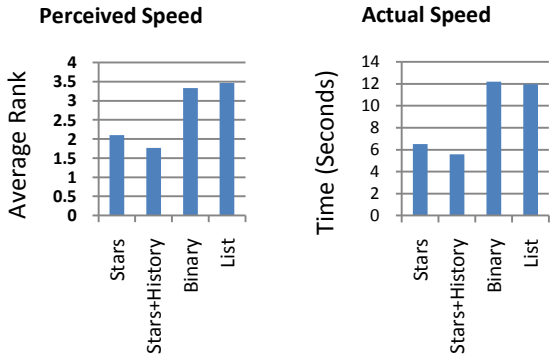

Figure 7. Average ranking of interfaces based on perceived and actual speeds (*N*=16). Note that the y-axes differ. Smaller numbers represent faster speed.

Other: We found no significant effect of interface on the mental demand ($\chi_3^2=4.50$, $W=0.094$) and fun to use ($\chi_3^2=2.24$, $W=0.047$). However, 7 participants ranked the Binary interface as the most fun to use.

Qualitative Analysis

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 higher speed with the Stars and the Stars+History interfaces based on their ease of use. Participants who thought the Stars+History was faster to use that the Stars interface, talked about how it helped them remember their previous ratings and calibrate their ratings as P14 said *"everything is in front of you… if you*

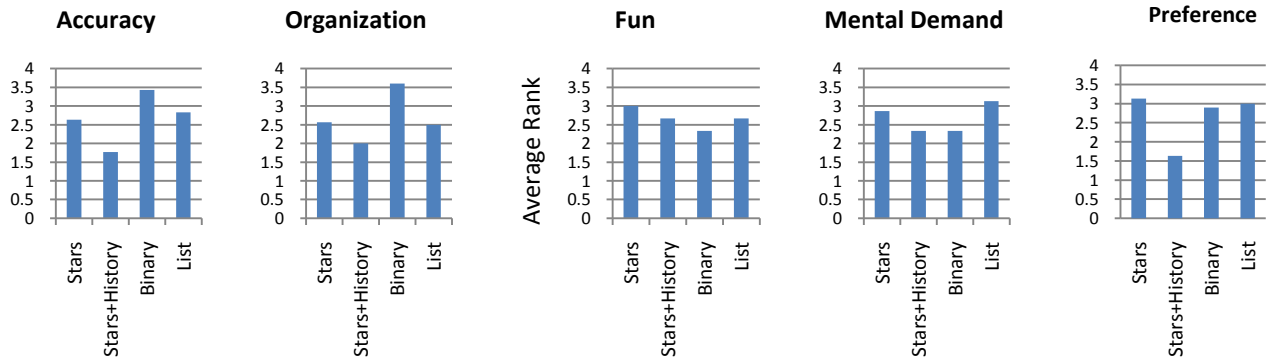

Figure 8. Average ranking of interfaces based on accuracy, suitability for organization, fun, mental demand, and overall preference (*N*=16). Shorter bars indicate better ranks.

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 Stars+History said that *"[With Stars] you don't compare, you just say something"*.

Accuracy: Some users felt that the extra information given in the Stars+History interface affected the level of accuracy they felt they should achieve. In reference to that interface 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 P7 said *"now that I see the history I do care… it allowed me to care about my accuracy*".

Interviews showed that 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 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 dragging a movie into the list. As mentioned earlier, we thought that having auto-scroll would have sacrificed generalisability to interaction with long lists of movies. 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 and 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.

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: P13 explained her preference for relative ranking over absolute rating, saying that *"you don't have to quantify anything. You just sort them."* Several participants indicated that they try to be consistent with their ratings and to remember their previous ratings, and that Stars+History facilitated the process. P4 compared the Stars and Stars+History 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 Stars+History] you just get reminded what the definition of a 7 and an 8 is."*

In regards to the poor ranking of the List interface (though not significant), two participants mentioned that with the List interface 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"* On the other hand, P9 preferred the list interface and mentioned that *"For Binary, I think for each comparison, while for the list, you have a sense of where it should be."* Moreover, several participants mentioned that sometimes it was hard to compare two movies when their merits were not comparable. For example, P10 said *"Sometimes they are not comparable … One is funny, one has a great story"*.

Suitability for Organization: Interviews showed that two main factors determine the suitability of interfaces for organization: first, the accuracy of the opinion captured by the interfaces and second, the preference of users for having a ranked list or having multiple categories. 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 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. Some others preferred to have categories, as in star interfaces. For example P7 said: *"The categorization helps me a lot more that 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."*

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 Stars+History was visually appealing"* (P12). These participants responded to this question based on the ease of use or low mental demand of interfaces. Some other users found the Binary interface fun to use and mentioned that the comparisons in the binary interface were *"almost like playing a game"* (P10) or *"like competition between movies"* (P9). Regarding the List interface, P10 mentioned that *"it was fun to see the order of movies in the list"*.

Overall Preference: Participants talked about various aspects of interfaces that influenced their preference. The 10 participants who preferred the Stars+History 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 Binary, talked about its simplicity and fun. The 2 participants who preferred the List interface talked about the fine granularity supported by it and their preference for ranking over rating. The only participant who preferred the Stars interface (P1) mention that 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

Our work raises several interesting issues with respect to designing rating interfaces. The results show that providing contextual information can support expressing opinion without sacrificing speed. The Stars+History interface

provided context while taking advantage of positive knowledge transfer. It did not require users to think about expressing their opinion in a conceptually different way (relative to the common Star interface), as did the Binary and List interfaces. In this study, the Stars+History interface tended to be faster and more accurate than the other interfaces (though not significantly); however, the speed-accuracy tradeoff can still be important for future interfaces. Most of our participants expressed that they care more about accuracy than speed of rating. However, the importance of speed may differ based on rating strategies; for example, when rating movies in batches the accumulated time becomes significant and users may consider speed a more important factor.

A general comment on the Binary and List interfaces was that people had difficulty comparing movies that were hard to compare. Based on the prototype walkthroughs and the formative study we knew that it would be easier to rank movies from different genres in different lists, however, it was not practical to ask the participants to provide a list of 20 movies from a single genre. It is important to provide relevant context to facilitate the cognitive process of expressing opinion. In the elaboration phase of reporting an opinion, people try to retrieve information relevant to the movie, including information about the movie itself and about the other related movies. Providing relevant context facilitates this process by reducing the burden on memory. However, irrelevant information or information that is hard to associate with the task imposes extra load by requiring an additional thought process to find connections between the provided context and the item being evaluated. In this study we did not consider genre or other aspects of movies to provide context, which can explain the absence of significant effect of interface on mental demand.

The overall poor performance of the ranking interfaces, namely the Binary and List interfaces can be attributed in part to familiarity bias. However, the Binary interface can perform poorly because every slip can be catastrophic; when the list goes out of order, every new insertion using binary search will be subject to error. The Binary interface is a focus-only interface, which makes it hard for the users to identify their errors when using it. This sensitivity led to the Binary interface being ranked poorly for accuracy and suitability for organization. On the other hand, many of the participants enjoyed using the Binary interface, which suggests that it can be used to collect bits of information about people's taste without necessarily using it as the primary interface for recording experiences.

A general concern about the List and Binary interfaces was scalability. For the List interface, the two strategies of using a small window into the list, and not supporting auto-scroll were devised to ensure that the overall interaction will not change significantly; however, they slowed down the participants. Moreover, information visualization techniques such as focus+context methods [6] can be used to facilitate the navigation of long lists. For the Binary interface, the number of comparisons grows very slowly. Moreover, one possibility for decreasing the number of clicks 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. One possibility is that, depending on the range of movies that participants had in their collections, the context provided by the interfaces (i.e. previously rated movies) had varying levels of usefulness. Another possibility is that the differences in the participants' preferences are related to their ability or willingness to take advantage of the previously rated movies through drawing connections between them and the movie being rated. Further studies are needed to investigate these hypotheses.

A limitation of our prototypes is 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 rating abstract concepts or items without a visual representation.

FUTURE WORK

We explored a small but important subspace of the design space of opinion measurement interfaces. Several dimensions of the design space are left unexplored. For example, in this study we did not investigate the precision of the n-star interfaces. The precision can be high enough to be perceived as continuous. Sliders have long been used for setting continuous properties and although, previous research has shown little difference in reliability and discrimination power of precise scales such as a 101-point with a 9-point scale [19], providing history as in Stars+History interface may enable users to benefit from the precision of those scales.

Another avenue for investigation is design of interfaces that help users make decisions about a product or service, through providing more informative representations of aggregated ratings, most commonly represented with the average rating in the current systems. Further studies are required to better understand the design space of opinion expression and representation interfaces. Future research on opinion expression interfaces should aim at providing relevant context, while fostering an understanding of how the context is relevant to the current item. Despite the underwhelming performance of the ranking interfaces, we believe they have merits that call for further investigation and design improvements. Specifically, the fun of using the Binary interface and the potential for organizing

experiences using the List interface are two of the avenues for future research.

SUMMARY AND CONCLUSION

We conducted a formative study to deepen our understanding of how and why people rate products and services on the Internet. Among many findings, we observed that all of the participants recalled relevant experiences and compared them with the experience or product being rated.

We focused on two of the dimensions of design space of opinion measurement interfaces: relativity (relative ranking vs. absolute rating) and amount of context (focus+context vs. focus only). We iterated over four conceptually different prototypes to explore the design space through a within-subject mixed-methods study. Overall, the participants preferred Stars+History interface that provided context for rating but did not require direct comparisons. Although people implicitly or explicitly compare movies to come up with a rating, this process turned out to be more complex than what we expected. It involves comparison with movies that are related based on criteria largely determined by the movie, and the movie rater's viewpoint and experiences.

In addition to the pervasive use of these interfaces for rating products and services, researchers in various disciplines use Likert scale and other simple interfaces to elicit people's opinion. Studying interfaces for measuring opinions is important: it will result in more accurate and internally valid subjective data. 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. *Computer Modeling and Simulation, Conference on*, 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.M., Tintarev, N., and Oliver, N. Rate it again: increasing recommendation accuracy by user re-rating. *Proceedings of the ACM conference on Recommender systems*, ACM (2009), 173–180.
4. Amatriain, X., Pujol, J., and Oliver, N. I Like It... I Like It Not: Evaluating User Ratings Noise in Recommender Systems. In G.-J. Houben, G. McCalla, F. Pianesi and M. Zancanaro, eds., *User Modeling, Adaptation, and Personalization*. Springer Berlin / Heidelberg, 2009, 247-258.
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? *Human Factors and Ergonomics Society Annual Meeting Proceedings 51*, (2007), 234-238.
6. Cockburn, A., Karlson, A., and Bederson, B.B. A review of overview+detail, zooming, and focus+context interfaces. *ACM Computing Surveys (CSUR) 41*, (2009), 2:1–2:31.
7. Cosley, D., Lam, S.K., Albert, I., Konstan, J.A., and Riedl, J. Is seeing believing?: how recommender system interfaces affect users' opinions. ACM (2003), 585-592.
8. Eagly, A.H. and Chaiken, S. *The Psychology of Attitudes*. Orlando, FL, US: Harcourt Brace Jovanovich College Publishers, 1993.
9. Gawronski, B. Editorial: Attitudes can be Measured! But What is an Attitude? *Social Cognition 25*, 5 (2007), 573-581.
10. Harper, F.M., Li, X., Chen, Y., and Konstan, J.A. An Economic Model of User Rating in an Online Recommender System. In L. Ardissono, P. Brna and A. Mitrovic, eds., *User Modeling 2005*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, 307-316.
11. Herlocker, J.L., Konstan, J.A., Borchers, A., and Riedl, J. An algorithmic framework for performing collaborative filtering. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, (1999), 230–237.
12. Ivanov, A., Erickson, T., and Cyr, D. Plot-polling: Collaborative Knowledge Visualization for Online Discussions. *Conference on Information Visualisation, International*, 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. *International Journal 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. McNemar, Q. Opinion-attitude methodology. *Psychological Bulletin 43*, 4 (1946), 289-374.
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. *CHI '06 extended abstracts on Human factors in computing systems*, (2006), 1181–1186.
18. 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.
19. Van Schaik, P. and Ling, J. Design parameters of rating scales for web sites. *ACM Transactions on Computer-Human Interaction 14*, 1 (2007), 4-es.
20. Thurstone, L.L. Attitudes Can Be Measured. *The American Journal of Sociology 33*, 4 (1928), 529-554.
21. Turnbull, D. Rating, voting & ranking: designing for collaboration & consensus. *CHI '07 extended abstracts on Human factors in computing systems*, ACM (2007), 2705–2710.
22. Ye, Q., Law, R., and Gu, B. The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management 28*, 1 (2009), 180-182.