Is Seeing Believing? How Recommender Interfaces Affect Users' Opinions

Dan Cosley, Shyong K. Lam, Istvan Albert, Joseph A. Konstan, John Riedl
Department of Computer Science and Engineering
University of Minnesota, Minnesota, MN 55455 USA
+1 612 624-8372
{cosley, lam, ialbert, konstan, riedl}@cs.umn.edu

ABSTRACT

Recommender systems use people's opinions about items in an information domain to help people choose other items. These systems have succeeded in domains as diverse as movies, news articles, Web pages, and wines. The psychological literature on conformity suggests that in the course of helping people make choices, these systems probably affect users' opinions of the items. If opinions are influenced by recommendations, they might be less valuable for making recommendations for other users. Further, manipulators who seek to make the system generate artificially high or low recommendations might benefit if their efforts influence users to change the opinions they contribute to the recommender. We study two aspects of recommender system interfaces that may affect users' opinions: the rating scale and the display of predictions at the time users rate items. We find that users rate fairly consistently across rating scales. Users can be manipulated, though, tending to rate toward the prediction the system shows, whether the prediction is accurate or not. However, users can detect systems that manipulate predictions. We discuss how designers of recommender systems might react to these findings.

Keywords

recommender systems, collaborative filtering, personalization, persuasive computing, e-commerce, conformity

INTRODUCTION

Humans' ability to locate the information they desire grows more slowly than the rate at which new information becomes available. Recommender systems are one tool to help bridge this gap. These systems use people's opinions about items in an information domain in order to help people make decisions about which other items to consume. For example, Amazon.com allows a user to rate books, then suggests other books the user might like based on those ratings.

CHI 2003, April 5–10, 2003, Ft. Lauderdale, Florida, USA.

Recommender systems can be a valuable competitive advantage to retail companies, especially in e-commerce. A system that produces good recommendations can inspire trust in the company and help users find products they truly want. At the same time, an engaging interface for collecting recommendations allows the company to gather preference information from its customers and tailor offerings to each customer. Both the company and its customers stand to benefit.

Few researchers have investigated the effect of interfaces on the use of recommendations. Herlocker et al. studied how explaining recommendations can convince users to trust the system [9], while Swearingen and Sinha find that users trust systems that recommend items the users know they like [16]. Instead, most research in recommender systems has focused on discovering good algorithms (e.g., [3, 8, 12, 14, 15]). This lack of attention to the interface poses a danger to users. In this paper, we focus on how the rating interface may affect both users' opinions and their ability to express them.

Recommendations may influence users' ratings

Recommender systems generally provide information about the items they recommend. This may include item descriptions, reviews written by other users or professional critics, average user ratings, or predicted personalized ratings for the given user. Even the fact that the item is recommended provides information—the system thinks the user will like this item. Recommender systems often provide a way for users to rate an item when it is recommended. Figure 1 shows the MovieLens interface, which is like many others in including predictions and a ratings interface on the same screen.

Showing information about an item at the time a user rates it might affect the user's opinion, leading to three potential problems. First, the altered opinion might provide the recommender with less accurate preference information, leading to less accurate predictions in the future.

Second, the altered opinions might make it hard to evaluate the quality of a system's recommendations. A system whose interface steers users toward its predictions might score better on accuracy metrics than a system with a more neutral interface, even though the second system might produce more useful recommendations.

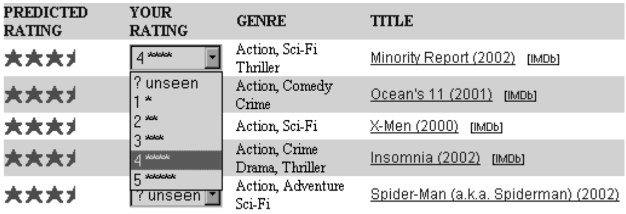

Figure 1: The interface for viewing predictions and rating movies in MovieLens. Predictions are displayed right next to the rating interface.

Third, unscrupulous agents might take advantage of this effect to amplify false opinions they inject into the system. Such opinions might be artificially inflated, leading to unusually positive recommendations which may in turn induce unusually positive ratings from other users. An agent might also attempt to prevent an item from being recommended by giving it falsely low opinions. Those who rate the item later may be swayed to give lower ratings as well. The more a system's interface influences users' opinions, the more effective and tempting such shilling attacks would be—particularly if users cannot tell that the predictions are being manipulated.

Mapping opinions to ratings is complex

Many recommenders represent a user's opinion about an item as a single number on a rating scale. These scales vary widely in their granularity. E-commerce systems often use purchase decisions as a proxy for ratings, resulting in either a unary scale (bought items are "liked", others unknown) or a binary scale (bought items "liked", unbought items disliked). Other systems ask users to rate on a 1-to-5-star Likert-style scale. Launch.com allows users to rate songs on a scale from 0 to 100, plus a control we call the "Britney Spears button" that allows the user to never hear a particular song again. The Jester joke-rating system lets users click a continuous bar, generating ratings from -10 to $+10$ [8]. Moviecritic.com used a 14-point scale. Tivo asks for ratings from -3 to $+3$.

Who's right? Or, more generally, what qualities should a rating scale possess? Ideally, a rating scale should allow users to express their opinions in a meaningful way without too much effort. This can be tricky, since opinions can be complex. Consider a user of a research paper recommender trying to assign a rating to a paper. Her opinion depends on the importance of the topic, the quality and originality of the research, the quality of the writing, the relevance of the work to her research, her mood when reading the paper, and so on.

Recommender systems work well despite asking users to map complex opinions to a number between 1 and N. We suspect that identifying good values for N will help recommender systems work better. Presumably, N should be high enough so that users can create the correct number of categories for them to distinguish between levels of liking, but not so high that users can't make judgments between the categories. The scale should also allow the system to make accurate predictions. Finally, the user should be able to make sensible evaluations of how much to trust those predictions and the recommender as a whole.

Intuitively, a fine-grained rating scale seems most likely to have these properties. MovieLens users' number one request is to rate movies on a half-star scale, while Swearingen and Sinha find that users prefer the continuous feel of the Jester-style interface [16]. Does finer granularity lead to better recommendations and happier users? Will users be able to make fine distinctions between levels of liking? Or will the additional expressiveness just produce noise?

Our Contributions

To the best of our knowledge, no one has studied how recommendations affect users' opinions of the items recommended. The extent to which these effects occur in practice may have dramatic importance for the design of recommender system interfaces and their practical implementation. We believe it is important to characterize and publish these effects so recommender systems designers and users can plan for them.

We conduct three experiments with a total of 536 users in order to answer the following questions:

- How consistent are users when re-rating items?
- What do users want in a rating scale?
- How do different rating scales affect users' ratings?
- Does the rating scale affect prediction accuracy of common collaborative filtering algorithms?
- How does showing predictions affect users' ability to re-rate items consistently?
- What happens if the system shows deliberately incorrect predictions when users re-rate movies?
- Can the system make a user rate a "bad" movie "good"?
- What happens if the recommender shows deliberately incorrect predictions for movies not yet rated?
- Do users notice when predictions are manipulated?

The remainder of this paper addresses these questions. We first survey related work and establish a theoretical basis for our research questions. We then outline three experiments we performed to address the questions, and tackle the questions one by one. Finally, we discuss the implications of the answers for recommender system designers and researchers.

RELATED WORK

Recommender Systems

Recommender systems use a number of strategies for modeling users. A common model is for *users* to assign *ratings* to *items*. When a target user wants recommendations, the system calculates *predictions*, estimates of how the target user would rate the items. It then typically recommends items with high predicted ratings.

Content-based filtering and collaborative filtering (CF) are two broad classes of strategies for computing predictions. Content-based systems often build a profile of keywords from items users like and recommend new items which match the profile. This strategy works well in text domains but does not work well when the content is hard to analyze.

Instead of finding similarity between the content of items, CF systems find similar *users* to a target user by comparing users' opinions of items. Many common CF systems compute similarity between users by comparing vectors of ratings using Pearson correlation, cosine similarity, or other distance metrics. These systems make predictions by computing a weighted average of the votes of similar users.

Collaborative filtering works in a number of domains. Resnick et al. used this approach to filter Usenet news in the GroupLens system [13]. Shardanand and Maes built Ringo, a music recommender [15], while Hill et al. built an early recommender for movies [10]. A number of other systems have been built, and CF is a widely used strategy for recommending items in e-commerce.

We introduce the basic ideas of collaborative filtering because our experiments use MovieLens, a CF recommender system for movies. However, we believe that most of our results apply to recommender systems in general.

Ratings Consistency

Hill et al. [10] asked users to re-rate movies they had rated six weeks earlier. The 19 users who responded had a strong correlation (0.83) between their earlier and later ratings. Pennock et al. assume that users give ratings from a Gaussian probability distribution in their personality diagnosis algorithm [12], explicitly recognizing that people may rate the same item differently at different times.

Our work extends Hill et al. by measuring how consistently users re-rate items on different scales, as well as how seeing predictions when rating affects users' consistency.

Tricking Recommenders And Influencing Users

Every so often, an angry MovieLens user complains that shills are giving high ratings to bad movies in an attempt to deceive the system. Although we have seen little evidence of such attacks in MovieLens[1], such attacks are both plausible and detrimental to users.

Dellarocas outlines several possible attacks against recommender systems along with a strategy for minimizing the impact of these attacks [4]. Domingos and Richardson explore how to target marketing to users of a recommender system by looking for members who are influential in generating recommendations for other users [5].

In this paper, we look at how much impact a successful attack might have on users' decisions by seeing how their ratings change when predictions are artificially manipulated.

[1] Newly added movies do often receive higher ratings at first. We believe this is because the users who are most apt to like a movie will tend to be the first users to see and thus rate the movie.

Design of Rating Scales

The choice of rating scales is a major concern in survey design. Friedman and Amoo examine several aspects of designing Likert-style scales, including the labels associated with each choice, question interpretation, rating scale balance, ordering of choices, and number of choices [6]. Amoo and Friedman also show that changing a scale's range (e.g., from $-5..5$ to $0..10$) can affect the distribution of responses [1]. Garland suggests that excluding a middle choice can reduce respondents' bias toward providing positive replies, but that doing so can produce distorted results [7].

This work suggests that the effectiveness of rating scales is domain-dependent. We empirically investigate the effect several different rating scales have on user satisfaction and prediction accuracy in the area of recommender systems.

Conformity and Persuasive Computing

Psychologists have studied how other people's opinions may affect one's own. The classic conformity study by Asch asked subjects to compare three lines to a reference line and to choose the line with the same length as the reference line [2]. Each subject performed the task 12 times. The answer was easy to see in all cases. However, subjects made these choices while sitting with a small group of confederates of the experimenter who deliberately made wrong choices. One-third of all trials ended with the subject making an incorrect choice, and most subjects made at least one wrong choice. Several factors seem to contribute to conformity, including the desire to fit in with group norms and the fact that one is receiving information from the opinions of others (though in the experiment, the information was incorrect).

Of course, computers are not people. Will the conformity effect appear based on information provided by a computer, rather than another person? Will people feel an urge to "fit in" with the computer's opinions on movies? Studies in the area of persuasive computing suggest that the answer is "yes". Nass and Moon survey a number of their experiments in [11]. For instance, people react toward interfaces portraying a gender or ethnicity for the computer much as they do to actual people of that gender or ethnicity. People are also "polite" to computers, being less likely to give critical evaluations of a computer's performance if the same computer asks for the evaluation, compared to a different computer asking for the evaluation of the first.

The literature on conformity and persuasive computing provides a theoretical framework for our research questions.

EXPERIMENTS

We conducted a series of three experiments with the MovieLens recommender system, which uses collaborative filtering to make recommendations. MovieLens has about 70000 users, 5600 movies, and over 7 million ratings.

Each experiment asked users to rate a set of movies. For each movie, we chose a baseline that was either the user's prior rating for that movie (if they had rated it before) or our best prediction for their rating on the movie (if they hadn't). We compared the ratings given during the experiment to the baseline. In some cases, we also used the baseline when computing a prediction to show users while they rated a given item. We describe these experiments in more detail below.

RE-RATE: Re-rating movies while showing "predictions"

For each user in this experiment, we randomly selected 40 movies that they had previously rated at 2, 3, or 4 stars. We limited the movies to ratings in the middle of our 5-star scale so the ratings could change either positively or negatively. We asked users to re-rate the movies, recording each $\{$original rating, re-rating$\}$ pair.

The system presented four screens of 10 movies. No predictions were shown on one screen. The other three screens showed 10 movies with a prediction equal to the user's original rating, 10 with a prediction one star above the original rating, and 10 with a prediction one star below the original rating. Balancing the manipulation leaves the mean prediction unchanged; only the variance increases. The 30 movies with predictions were randomly distributed across the three screens in an effort to disguise the manipulation.

UNRATED: Manipulating predictions for unrated movies

This experiment is similar to RE-RATE. For each user we selected 48 movies they had *not* rated. MovieLens normally rounds predictions to the nearest half-star; since we wanted the experiment to be a similar as possible to RE-RATE, we chose movies for which our best prediction was within 0.25 stars of 2, 3, or 4 stars. Users were divided into an experimental group and a control group. For the experimental group, we divided the movies into four sets: no prediction, actual prediction, prediction plus one star, and prediction minus one star. Movies were displayed similarly to RE-RATE. We compared experimental users' ratings to their actual predictions. The control group performed the same task except that all shown predictions were actual predictions. We surveyed both groups' satisfaction with the experimental predictions and MovieLens predictions in general.

SCALES: Re-rating movies on other rating scales

For each user, we chose 45 movies they had rated. We required that the user had rated at least seven movies with each of the five different ratings on the MovieLens rating scale. We randomly divided these movies into three sets of 15 and asked users to rate each set on one of three rating scales:

- **Binary**: Thumbs up or thumbs down.
- **No-zero**: A scale from -3 to $+3$ with no zero.
- **Half-star**: A 0.5 to 5 star scale in half star increments.

We used different movies on each scale, rather than the same set of 15 movies on all three scales, because we were afraid that rating the same movie several times would bias the way users mapped their ratings to each scale. We looked at how new ratings mapped to original ratings. A follow-up survey a month later asked how well users liked each scale.

User Selection

MovieLens provides infrastructure for conducting experiments with real users. If a user logs in while an experiment is active and the user is qualified for the experiment, the user sees a link on the main MovieLens page, asking him if he would like to participate. If he clicks the link, the system presents a consent form. Users who consent to the experiment are randomly assigned to an experimental group.

This process makes it easy to conduct experiments with real users of MovieLens. However, it does bias the selection process toward users who log in more often and who are willing to participate in experiments.

In RE-RATE, 212 users re-rated 7574 movies. A total of 274 users participated in UNRATED, with the experimental group of 137 users giving a total of 1599 new ratings. Finally, 77 users re-rated 2795 movies on the three scales of SCALES. Each user participated in only one experiment.

ANSWERING OUR QUESTIONS

We now set out to answer the questions we posed earlier.

How consistent are users when re-rating items?

Hill et al. found that the correlation between 19 users' original ratings of a set of movies and the same users' ratings on the same movies six weeks later was 0.83 [10]. We were interested in seeing whether this correlation still held when ratings had been given months or even years before, as users' tastes and opinions can change.

We looked at how consistently users re-rated movies for which we showed no prediction in RE-RATE. 212 users provided a total of 1892 ratings in this portion of the experiment. Users re-rated at their original rating 60% of the time, below it 20%, and above it 20% of the time. The mean new rating was within 0.01 stars of the mean original rating. The correlation between original ratings and re-ratings was 0.70.

These data suggest that users are reasonably consistent when re-rating movies. Our correlation is lower than Hill et al.'s, but still quite strong. One possible explanation for the reduced correlation is that we limited the set of re-rated movies to those with original ratings of 2, 3, and 4, which allowed users to err in both directions.

What do users want in a rating scale?

Previous work has also suggested that users prefer finer-grained rating scales. To verify this, we conducted a follow-up survey on the users in SCALES. We asked users how well they liked rating on the binary, no-zero, and half-star scales, as well as on the original MovieLens scale. We also asked them to rank the four scales in order of preference. Of the 77 users, 26 responded to this survey. Users liked the half-star scale most (average satisfaction of 4.2), followed

	Binary	No-zero	Half-star
Total ratings	941	918	935
Original mean	0.589	0.605	0.576
New mean	0.705	0.641	0.579
Correlation	0.706	0.827	0.829

Table 1: User ratings on new scales versus original ratings. All scales were normalized to a zero-to-one range. Users rate significantly higher on both the binary and no-zero scales ($t(940) > 100, p < 0.01$; $t(918) = 2.60, p < 0.01$). Ratings correlate strongly on all scales.

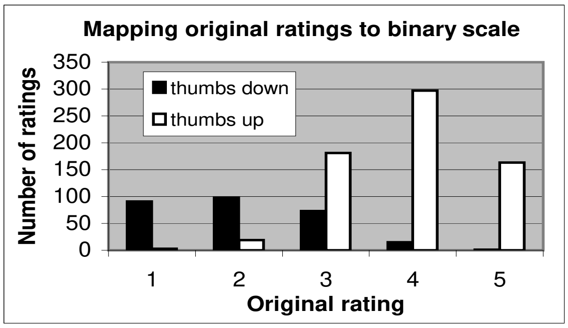

Figure 2: How users mapped original ratings to the binary scale. Original ratings of 1 and 2 are predominantly thumbs down, while higher ratings map to thumbs up.

by the original MovieLens scale (average 3.8), the no-zero scale (3.2), and the binary scale (2.2).

It appears that users do like a finer-grained scale the best. However, granularity is not the only factor, or else users would like the no-zero scale better than the original MovieLens scale. It may be that users were more comfortable with the familiar original scale. Another possibility is that the no-zero scale put too much emphasis on bad movies. One user said that the no-zero scale "[goes] too deep into ratings of bad movies and not deep enough into good movies. I don't go to see movies that I expect are really bad so I do not need a three point scale to rate these movies."

How do different rating scales affect users' ratings?

We now look at how the rating scale affects the way users map their opinions to ratings. For each scale, Table 1 shows how many ratings 77 users gave on the scale, the mean original rating for movies rated on that scale, the mean new rating, and the correlation between the old and the new ratings.

Users gave higher mean ratings on the binary and no-zero scales. However, new ratings correlate strongly with original ratings on all three scales, although less so for the binary scale. The shape of the distributions is also similar on both of the finer-grained scales.

Figure 2 shows how users mapped their original ratings to the binary scale. Users generally mapped original ratings of 1 and 2 to "thumbs down" and original ratings of 3, 4, and 5 to "thumbs up". The tendency to rate 3 as thumbs up on the binary scale explains most of the increase in average rating. Users seem to give borderline movies the benefit of the doubt when forced to rate on a coarse scale.

	Binary	No-zero	Half-star
Total predictions	591	616	657
Original MAE	0.201	0.209	0.223
New MAE	0.245	0.204	0.205

Table 2: Recommendation accuracy using ratings on new scales versus accuracy using original ratings. Scales were normalized to a zero-to-one range. MAE is significantly higher for the binary scale and lower for the half-star scale ($t(590) = 2.59, p < 0.01$; $t(656) = 2.43, p < 0.05$).

Does the rating scale affect prediction accuracy of common collaborative filtering algorithms?

We also looked at whether the scale makes any difference in the accuracy of collaborative filtering predictions. We used the "All But 1" protocol from Breese et al. [3]. In this protocol, we remove one rating from the entire dataset, make a prediction based on the remaining data, and compute the absolute error between the prediction and the rating left out. Averaging this absolute error over all items for which the system can make predictions gives the system's Mean Absolute Error (MAE) on the dataset. To make MAE comparable between different scales, we normalized each scale to a continuous zero-to-one scale.

Table 2 compares the MAE for predictions made on the new scales versus predictions made using the old scales. The number of predictions and original MAE differs for each scale because SCALES asked users to rate a different set of movies on each scale. Compared to a five-star scale, the MAE is worse for the binary scale, about the same with a six-point scale, and better for the ten-point scale.

We hesitate to draw conclusions from the MAE results, although the differences are statistically significant for the binary and half-star scales. The relatively small datasets collected during the experiment produced MAE values higher than one would normally expect in a collaborative filtering system. Also, the MAE using the original ratings was higher with the half-star scale than the others. Still, the MAE does fall as scale granularity increases. This might be because the size of some prediction errors will drop as the "quantum" of rating becomes smaller, bringing down the average error.

How does showing predictions affect users' ability to re-rate items consistently?

We now turn back to RE-RATE to see how showing predictions when users rate movies affects their ratings. We compare the movies users re-rated without seeing predictions with the movies they re-rated while seeing an accurate "prediction"—the user's original rating.

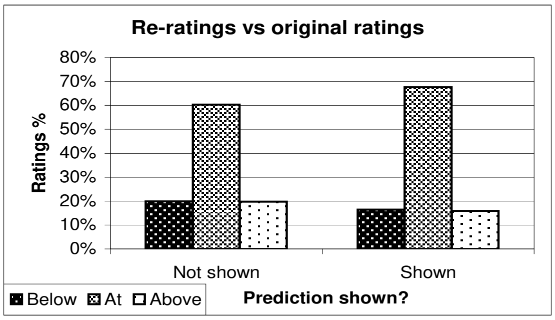

Figure 3: Percentage of re-ratings below, at, and above the original ratings, broken down by whether the original rating was shown as a prediction. Showing predictions causes users to rate significantly more often at their original rating ($\chi^2(2) = 21.8, p < 0.01$).

Figure 3 shows how often users re-rated movies below, at, and above their original rating, depending on whether they saw a prediction or not. Users' average re-rating was within 0.01 of their average original rating in both cases. However, users re-rated at their original rating significantly more often when they saw predictions.

One interpretation of the fact that users rate with lower variance when they see predictions is that it helps them remember their old rating, reducing "noise" in the re-ratings. A different interpretation is that the lower variance means we are actually influencing people's beliefs, convincing them to rate at the prediction shown by the system.

What happens if the system shows deliberately incorrect predictions when users re-rate movies?

To decide which interpretation was correct, we looked at re-ratings users gave to movies where the system altered the prediction one star above or below the user's original rating. In the RE-RATE experiment, users saw a total of 30 movies, 10 each in the accurate, up, and down conditions. The movies were randomly ordered to disguise the manipulation. Users gave a total of about 1900 ratings in each condition.

Figure 4 shows how often users re-rated movies below, at, and above their original rating when the system showed predictions that were one star below, at, and one star above the original rating. Users rate above or below their original rating more often and have higher or lower mean ratings when the system shows higher (mean $+0.14$) or lower (mean -0.16) predictions, compared to when the system shows accurate predictions. These differences were statistically significant.

Can the system make a user rate a "bad" movie "good"?

We wondered if there might be "sticking points" in the rating scale which would be hard to influence people to cross. Remember that in SCALES, users mapped 1- and 2-star ratings to "thumbs down" and 3-to-5-star ratings to "thumbs up". Perhaps the system can get users to rate 2-star movies lower and 3-star movies higher, but not the reverse.

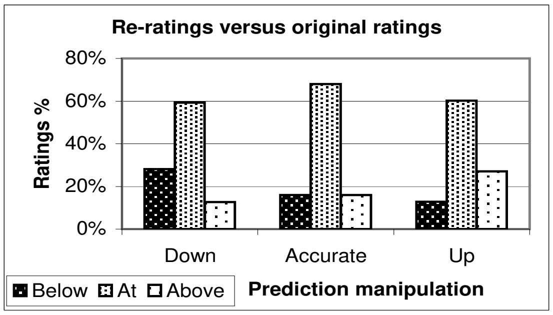

Figure 4: Percentage of re-ratings below, at, and above the original ratings, broken down by how the prediction was manipulated. Showing predictions altered downward or upward causes users to rate significantly lower or higher ($\chi^2(4) = 261, p < 0.01$).

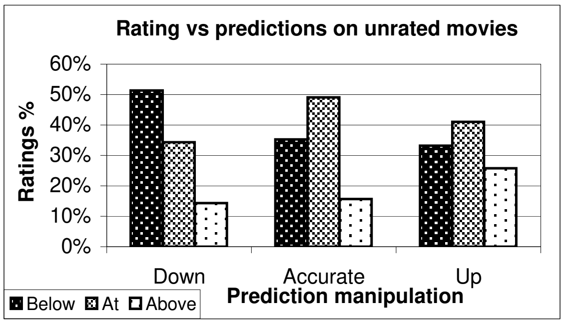

Figure 5: Percentage of re-ratings below, at, and above predictions when rating previously unrated movies, depending on how the prediction was manipulated. Showing predictions altered downward or upward causes users to rate significantly lower or higher than they do when accurate predictions are shown ($\chi^2(4) = 48.5, p < 0.01$).

This was not the case. No matter whether the original rating was 2, 3, or 4 stars, the effect on new ratings in RE-RATE was the same. This suggests that a recommender system can influence users to move from a negative to a positive rating.

What happens if the recommender shows deliberately incorrect predictions for movies not yet rated?

Since the user may never have chosen a star rating for movies he has not yet seen, we expect that showing predictions when users rate movies for the first time will have an even stronger effect. We turn to UNRATED to explore this question. UNRATED was very similar to RE-RATE, except instead of using movies the user had rated, we used movies the user had not yet rated. Also, since we did not have an original rating to use as a baseline, we used the prediction computed by the Net Perceptions recommender engine as the baseline.

Figure 5 shows how often users rated below, at, and above their predictions when shown a prediction that was lowered by one star, accurate, or raised by one star. Users gave about 400 ratings for each condition. Users rated below the actual prediction in all three cases; however, the pattern of users rat-

	Experiment				MovieLens			
	Accurate		Useful		Accurate		Useful	
	C	E	C	E	C	E	C	E
Excellent	13	8	8	4	22	14	30	27
Good	63	48	56	38	84	73	77	62
Fair	31	40	35	44	21	28	16	24
Poor	9	18	11	14	0	4	0	7
Awful	0	0	2	2	0	0	1	1
$\chi^2(1)$	6.32		5.44		3.90		6.21	

Table 3: User opinions of the accuracy and usefulness of experimental recommendations and MovieLens recommendations in general, expressed as percentages. Taking "Excellent" and "Good" as positive and "Fair" or below as negative, the control group (C) has significantly more positive evaluations than the experimental group (E) on all four questions ($p < 0.05$).

ing toward shown the prediction is clear. Compared to when actual predictions were shown, users' mean rating was 0.15 stars higher when seeing inflated predictions and 0.23 stars lower when seeing lowered predictions. These differences were statistically significant. Users also rated at their prediction significantly more often when predictions were shown compared to when they were not ($\chi^2(2) = 6.77, p < 0.05$).

Do users notice when predictions are manipulated?

It appears that showing predictions on unrated movies does lead users to rate in the direction of the prediction. This is bad news for users and good news for shills—unless users can detect the manipulation.

In UNRATED, a control group performed exactly the same rating task as the experimental group, except that whenever the system showed a prediction, it showed an actual prediction. We then asked both groups to complete a survey about the accuracy and usefulness of the recommendations they received during the experiment and the accuracy and usefulness of MovieLens recommendations in general.

Table 3 shows users' responses. The control group expressed significantly more satisfaction than the experimental group. We believe that the experimental users sensed that predictions were inaccurate and that this inaccuracy led to an overall decrease in liking of MovieLens.

PUTTING IT ALL TOGETHER

We have shown that a recommender's predictions do influence the way that users rate movies. Does this mean that we are actually changing the user's opinion, or are we just affecting their expression of their opinion as a rating?

We believe that showing predictions does change people's opinions. In addition to our findings, Asch reports that some people actually believed a longer line is shorter when they gave wrong answers. Others decided that their judgment is wrong, and adopt the group's judgment instead [2].

We don't know how long the change in opinion lasts. It would be more interesting if the change lasts longer than just for the moment of rating. We believe that the change will be lasting—that, once a person has rated a movie 4 stars, they will tend to think of it as a 4-star movie in the future.

Whether the system influences a user's opinion or not, and no matter how long the change in opinion lasts, the effect on ratings makes a difference from the practical standpoint of making recommendations. If a system receives higher ratings for *Dude, Where's My Car?*, it will tend to calculate higher predictions for the movie and recommend *Dude* to more users. This increases the value of manipulating the recommender, at least to the manipulator.

We have long believed that recommender systems are self-correcting: that, if artificially high ratings are given for an item, other users will give true ratings for that item that will cause it to not be recommended any more. These results suggest that self-correction may be reduced by the influence the manipulated predictions have on later ratings.

However, it is not open season to manipulate users. Even though accurate predictions have some error in them, users can detect the additional error in the manipulated predictions. Turpin and Hersh found that users of two search engines, one much better than the other, showed equal satisfaction with the systems [17], suggesting that users are not very sensitive to differences in search engine quality. By contrast, our results show that at least experienced users of a recommender system are sensitive to quality.

What does it mean for designers?

MovieLens users sometimes ask for the ability to hide predictions when they are rating items. They rightly suspected that seeing the prediction could influence their rating. To make users happy and to learn their preferences accurately, designers should accommodate them. It is convenient for users to allow them to rate an item whenever the system shows the item. Interface designers should consider designs that allow users to concentrate on rating while ignoring the prediction.

Users also prefer finer-grained rating scales. Since they seem to have no adverse effect on prediction accuracy, this too seems like a good idea. In fact, since user ratings correlate very well between scales, a designer might choose to allow users to rate on any scale they wish, computing recommendations using normalized scores.

Designers should take care that the scale allows users to make *meaningful* distinctions. For instance, users may not need to distinguish between degrees of badness. Launch.com's interface allows a 0-to-100 rating plus a "never play this again" option that fits this model. Such a system must make clear that the rating scale is measuring levels of goodness, however. Several authors of this paper use Launch. Sometimes Launch plays a song with the explanation "You rated this song"—even if the rating was a 10. We thought this was a low rating, but the system apparently did not.

Finally, it appears that users are sensitive to the manipulation of predictions. Their sensitivity to manipulation suggests that they will also be sensitive to inaccuracy, so it is important to choose a good algorithm. Just how good it must be depends on how sensitive users are to inaccuracy—which is still an open question—but we know that two-thirds of the items being off by one star was too much.

CONCLUSION AND FUTURE WORK

We found that showing predictions when users rate movies changes their ratings, although we don't know how long this change of opinion lasts. Would future "uninfluenced" re-ratings show the same bias? Also, does it matter what sort of item the system recommends? Users may rate movies less carefully than they might rate, say, a computer monitor. If this is the case, then users might be less influenced by the predictions of a computer hardware recommender.

An intriguing question is whether manipulating predictions will affect users' opinions of movies they haven't seen yet. Presumably, users will be more likely to go see movies with higher recommendations. Do users who see movies after receiving an artificial positive recommendation like the movies more than those who saw an artificial negative one?

We also saw that user satisfaction suffered when we manipulated ratings, probably due to lower accuracy. It would be interesting to see just how sensitive users are to inaccuracy, and whether they react differently to manipulation (known or unknown) vs. other forms of inaccuracy.

Recommender systems designers and researchers have primarily focused on delivering accurate recommendations. Much of the accuracy problem has been solved; well-tuned algorithms produce similar error patterns across a wide range of algorithmic approaches and data sets. Delivering these accurate predictions to users in a way that creates the best experience for them remains an open problem. The effect of presentation and interface is much less studied, and is likely the next area where significant improvements can be made.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation under grants DGE 95-54517, IIS 96-13960, IIS 97-34442, IIS 99-78717, and IIS 01-02229.

REFERENCES

1. T. Amoo and H. H. Friedman. Do numeric values influence subjects' responses to rating scales? *Journal of International Marketing and Marketing Research*, 26:41–46, February 2001.

2. S. E. Asch. Effects of group pressure upon the modification and distortion of judgements. In *Groups, Leadership, and Men*, pages 177–190, 1951.

3. J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proc. UAI-98*, pages 43–52, July 1998.

4. C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *Proc. 2nd ACM Conference on Electronic Commerce*, pages 150–157, Minneapolis, 2000.

5. P. Domingos and M. Richardson. Mining the network value of customers. In *Proc. SIGKDD 2001.*, pages 57–66, San Francisco, 2001. ACM Press.

6. H. H. Friedman and T. Amoo. Rating the rating scales. *Journal of Marketing Management*, 9(3):114–123, 1999.

7. R. Garland. The mid-point on a rating scale: Is it desirable? *Marketing Bulletin*, 2:66–70, 1991.

8. K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

9. J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *Proc. CSCW2000*, pages 241–250, 2000.

10. W. Hill, L. Stead, M. Rosenstein, and G. Furnas. Recommending and evaluating choices in a virtual community of use. In *Proc. CHI95*, pages 194–201, 1995.

11. C. Nass and Y. Moon. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 60(1):81–103, 2000.

12. D. M. Pennock, E. Horvitz, S. Lawrence, and C. L. Giles. Collaborative filtering by personality diagnosis: A hybrid memory and model-based approach. In *Proc. UAI-00*, pages 473–480, San Francisco, July 2000.

13. P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. GroupLens: An open architecture for collaborative filtering of netnews. In *Proc. CSCW94*, pages 175–186, 1994.

14. B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In *WWW10*, pages 285–295, 2001.

15. U. Shardanand and P. Maes. Social information filtering: Algorithms for automating "word of mouth". In *Proc. CHI95*, pages 210–217, 1995.

16. K. Swearingen and R. Sinha. Interaction design for recommender systems. In *Designing Interactive Systems (DIS2002)*, London, June 25–28 2002.

17. A. H. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *Proc. SIGIR2001*, pages 225–231, New Orleans, Sept. 9–13 2001.