# Confidence Displays and Training in Recommender Systems

**Sean M. McNee, Shyong K. Lam, Catherine Guetzlaff, Joseph A. Konstan, and John Riedl**

GroupLens Research Project

Department of Computer Science and Engineering

University of Minnesota, MN 55455 USA

{mcnee, lam, guetzlaf, konstan, riedl}@cs.umn.edu

**Abstract:** Recommender systems help users sort through vast quantities of information. Sometimes, however, users do not know if they can trust the recommendations they receive. Adding a confidence metric has the potential to improve user satisfaction and alter user behavior in a recommender system. We performed an experiment to measure the effects of a confidence display as a component of an existing collaborative filtering-based recommender system. Minimal training improved use of the confidence display compared to no training. Novice users were less likely to notice, understand, and use the confidence display than experienced users of the system. Providing training about a confidence display to experienced users greatly reduced user satisfaction in the recommender system. These results raise interesting issues and demonstrate subtle effects about how and when to train users when adding features to a system.

**Keywords:** Recommender Systems, Collaborative Filtering, Confidence Displays, Risk, Training

## 1 Introduction

People face the problem of information overload every day. As the number of web sites, books, magazines, research papers, and so on continue to rise, it is getting harder to keep up. In recent years, recommender systems have emerged to help people find relevant information.

Collaborative filtering (CF) is a family of algorithms commonly used in recommender systems. CF works by having users rate items (e.g., books or movies) and then matching users who express similar preferences. Thus, each user has a personalized "neighborhood" of similar users, all of whom have rated different items. The system recommends items that a user's neighbors like that the user has not seen. For example, MovieLens (www.movielens.org) is a CF-based movie recommender where users rate movies and receive recommendations on which movies to watch and which ones to avoid.

One strength of CF systems is that they can recommend items outside the user's usual content range. For example, a fan of science fiction may receive a recommendation for a particular art film because other science fiction fans liked it. This serendipity in recommendations is one of the strengths of collaborative filtering.

It is also one of its weaknesses. CF algorithms often recommend obscure items that are known only by a few loyal fans. Will users trust a system that recommends many obscure items? The popularity distribution of items in the recommender system compounds this problem. Though MovieLens has over 70,000 users and 6 million ratings, many of its 6,000 films have few ratings. If a fan of one of these obscure items is in a user's neighborhood, there is unlikely to be any second opinion. This can result in highly serendipitous recommendations, but also ones based on very little data.

This is not entirely bad, as people might find such recommendations to be useful. However, since the recommendation is based on little data, the recommendation is less likely to be correct. Believing these recommendations then poses a risk for the user. The less confident the system is, the greater the risk. By providing a confidence display for recommendations, the system could help users make better-informed decisions about whether to follow specific recommendations. This in turn could inspire trust in the system by making the recommendation process more transparent to the user.

Accordingly, we explore both the question of whether a confidence display should be added to recommender systems, and the equally important question of how to successfully add such a confidence display. We believe that these questions may have different answers for new versus experienced users since new users may be overwhelmed by a system and experienced users may have already developed sophisticated (and possibly inaccurate) mental models of how the system works.

For these reasons, we separately examined the reactions of new and experienced users. We also examined whether *training* improved users' reactions to the confidence display. Training could help experienced users adjust their mental models gently, while allowing new users to more easily navigate the interface and help them build more accurate models of how the system makes recommendations.

## 1.1 Contributions

We propose a basic confidence display to be used in recommender systems, and we answer the following questions about this display:

1. How does adding a confidence display change users' satisfaction with a recommender system?
2. How does adding a confidence display change users' behavior in a recommender system?
3. Do new and experienced users react differently to such a confidence display?
4. How does providing training on a confidence display in a recommender system affect user satisfaction and behavior?

The outline of the paper is as follows. First we explore related work in collaborative filtering and in confidence displays. Next, we propose and justify our confidence display. Then we present the design and results of our experiment. Finally, we discuss our findings, and state some implications for designers of recommender systems, which may prove useful to designers in general.

# 2 Related Work

Automated collaborative filtering was introduced in 1994 [Resnick 1994]. Herlocker et al. and Breese et al. provide overviews of various CF algorithms [Breese 1998, Herlocker 1999]. From the earliest work on CF systems, researchers have hoped to develop mathematical and statistical measures of confidence, but the large number of sources of errors has made this problem extremely difficult, if not intractable. While many new algorithms for computing recommendations have been developed (e.g., item-based CF [Sarwar 2001], Bayesian networks [Breese 1998] and factor analysis [Canny 2002]), none of these provide a statistical measure of the confidence of a recommendation.

Our work extends previous CF work by suggesting a basic confidence metric that can be used by any recommender system. We then experimentally evaluate a confidence display using this metric. This work can also be viewed as a way to increase trust in recommendation algorithms, an area investigated by Swearingen et al. [2002].

In the realm of confidence metrics and displays in interfaces, Carsewell et al. [1990] explored the interaction and interference between two display variables, such as for a recommendation and prediction metric. Herman et al. performed a study on fighter pilots' displays showing confidence estimates on enemy locations [Herman 1971]. Confidence metrics have also been explored for boosting algorithms in machine learning [Schapire 1999]. MovieCritic (www.moviecritic.com, no longer active) had a bull's-eye UI graphic to indicate a level of confidence in their movie recommendations (though it did not reveal the metric behind the display). Finally, Herlocker [Herlocker 2000] explored several different interfaces for explaining collaborative filtering recommendations.

We extend this work by studying how a confidence display in a recommender system affects users' opinion of and behavior in the system. Once we have seen how confidence affects users, we can revisit this work to develop more sophisticated confidence metrics and better ways to display them.

Mack et al. [1983] found that learning to use complex user interfaces can be extremely difficult for users. They proposed various training methods to help users learn the system. Since then, McGrenere et al. [2002] have shown that complex interfaces are still difficult to learn. Carroll [1990] has proposed that a simpler "learning interface" should be used for new users. Windows XP from Microsoft, for example, now includes training for interface elements including popup windows and wizards to guide users through unfamiliar parts of the user interface.

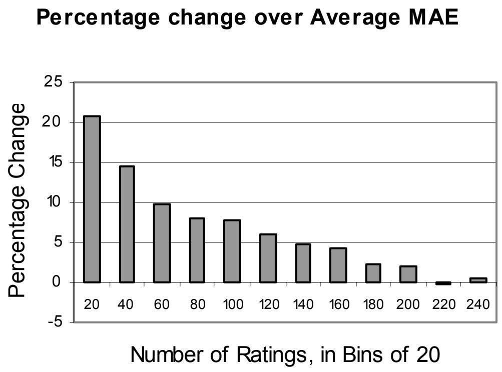


**Figure 1:** Percentage above average MAE in MovieLens recommendations, grouped by how many ratings the system has for each item recommended. The y-axis represents the percentage increase over the average, while the x-axis shows the number of ratings each movie had, binned into groups of 20. Movies with few ratings have higher error.

We investigate empirically how timing and training affects users' reactions to the introduction of a new interface element.

## 3 Confidence in Recommenders

There are many possible indicators of low confidence that could be used to generate a metric. A user-item CF algorithm could base a metric on the strength of the user's neighborhood; an item-item algorithm could use a measure of item similarity; and an SVD or factor analysis algorithm might consider the outlier items that complicate the reduction. Instead of focusing on one particular algorithm, however, we looked for a metric that all CF-based systems could employ.

Ratings are at the core of all CF-based systems. Whether explicitly entered or implicitly gathered, the more ratings the system has for a user, the better the system knows the user. Similarly, the more ratings for an item the system has, the better it knows that item. Thus, for an item with few ratings, it is plausible that a recommender system could make less accurate predictions.

Figure 1 shows that movie predictions in MovieLens for movies with many ratings are more accurate than those with few ratings. Specifically, the graph shows changes in mean absolute error (MAE) averages for recommendations, grouped by the number of ratings the system has for the movie being recommended. The movies are grouped into bins of 20: 1-20 ratings, 21-40, etc.

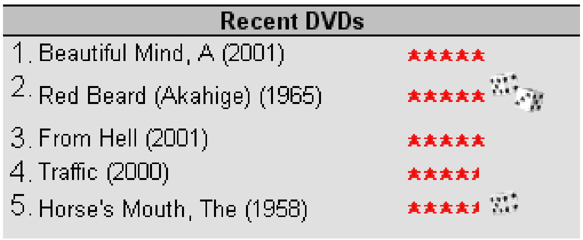

Recent DVDs
1. Beautiful Mind, A (2001) ★★★★★
2. Red Beard (Akahige) (1965) ★★★★★
3. From Hell (2001) ★★★★★
4. Traffic (2000) ★★★★½
5. Horse's Mouth, The (1958) ★★★★½

**Figure 2:** Dice icons indicate that The Horse's Mouth is risky, and that Akahige is very risky, because the system has few ratings for these items.

This graph suggests that the number of ratings an item has can be used as a simple, non-personalized measure of confidence for predictions in a recommender system. This metric is extremely simple to calculate and can be easily implemented in any recommender system, so we chose it as our confidence metric for the experiments in this paper.

Figure 2 shows our user interface display for confidence. Movies recommended with few ratings were considered "risky" movies. We chose to use dice as our visual indication of risk because they connote chance and uncertainty. After reviewing the raised tail at left end of Figure 1, we decided that movies with fewer than 40 ratings were considered very risky and were marked with two dice. Movies with 41 to 80 ratings were marked with one die. Currently in MovieLens, 29% of the movies would be considered risky, with two-thirds of these receiving two dice. Clicking on the dice takes users to a page briefly explaining how MovieLens determines predictions, why the prediction was risky, and how many ratings the movie had.

We realize that both the confidence metric and the corresponding display presented here are simple. Since we knew from surveys of MovieLens users that many of them desire some form of confidence display, we were more interested in knowing how adding a confidence display would affect users in the system than in developing the best possible metric. Thus, we chose to implement the simplest metric we could find that correlated well with actual error and focused our experiments on users' reactions.

### 3.1 Training

Users often ignore individual features as the number of features increases [McGrenere 2000]. Ideally, we can use affordances to help users understand what a feature does and how it is useful just by seeing it. When designed correctly, many features of a user interface should appear intuitive and facilitate easy and expressive communication between the user and the system.

When affordances are insufficient, designers can turn to other tactics. Shared social contexts and metaphors, encouragement of discovery, and training are all ways designers can get users to understand and use features of the interface. While training is heavy-handed, it is also direct [Mack 1983].

Training does not need to be the electronic equivalent of a three-day intensive seminar. A pointer is often all that is needed to make the feature apparent, especially for a secondary interface element such as our confidence display. In our experiments, user training consisted of a few sentences of text and one image. We hypothesize that even minimal training will increase users' ability to use and appreciate the confidence display.

# 4 Experiment

|  | Control | No Train | Train | *Total* |
|---|---|---|---|---|
| New Users | 37 | 32 | 44 | *113* |
| Experienced | -- | 57 | 53 | *110* |

**Table 1:** Number of users in each group. Users were divided into new and experienced MovieLens users.

## 4.1 Design

We performed an online study using both new and experienced MovieLens users. Table 1 shows how many users participated in each of the five groups. Users were asked to participate in a "user interface" experiment, told to use the system as they normally would, and that they would receive a survey to complete at the end of the three-week study.

When a new user consented to joining the experiment, she was randomly placed into one of three groups: a control group that did not receive dice icons, an experimental group that received dice icons without training, or an experimental group that received dice icons and the page of training on their use. The training page consisted of the following text and an image similar to Figure 2:

> "Take a look at the following fake recommendations. Notice that some of the recommendations are denoted with dice icons. This indicates movie recommendations that can be considered 'risky'. While using MovieLens you can click on the dice icons for an explanation of why that recommendation is risky."

Experienced users of MovieLens, after consenting to join the experiment, were randomly placed into one of two groups: an experimental group that received dice icons without training, and an experimental group that received dice icons and the same page of training as the new user experimental group.

## 4.2 Survey

After the experiment ended, we asked users the following questions:

1. How often do you think MovieLens predictions are correct?
2. How effectively do you think you can use the system to choose the right movie to see?
3. How happy have you been with your MovieLens experience?
4. Did you notice the dice icons?
5. Did you know what the dice icons meant?
6. How valuable were the dice icons in helping you make your movie selection?
7. Did you avoid those movies with the dice icon
8. Did you check other sources for information about movies with the dice icons before making a selection?
9. Were you more likely to check other sources for information about movies with the dice icons than for those without?

Finally, we asked users for any extra comments. The new-user control group received a slightly different version: they were not asked questions 4 and 5, and they were then exposed to the confidence display as a possible addition to MovieLens and then were asked hypothetical variants of questions 6 through 9.

## 4.3 Task-Based Study

Experienced users were also asked to perform three movie selection tasks using MovieLens with the confidence display. Each task presented one of three scenarios that we felt represented varying levels of risk in the movie selection process.

The first scenario, chosen to be "risk neutral", was to have the user choose a movie to watch by herself that evening. The second scenario was a "risk averse" scenario requiring the user to select a movie to watch with a distant, but important, family member. The third scenario was a "risk seeking" scenario in which the user gets together with close friends often to watch a wide variety of movies, and this week it is the user's turn to choose the movie. We did not explicitly tell users that the scenarios were risk neutral, risk averse, or risk seeking.

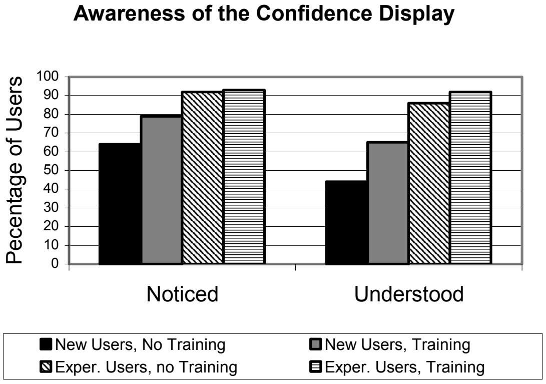


**Figure 4:** Differences in awareness of the confidence display, in percentages. Experienced users had more awareness and understanding of the display. Training increased awareness and understanding for new users.

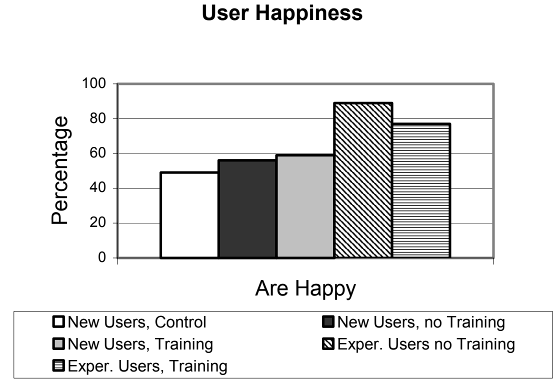


**Figure 5:** User happiness with MovieLens, in percentages. Experienced users are happier than new users, and training affects happiness in both groups.

The task required the user to enter the movie title of their selection into a text box on the survey. All experienced users were asked to complete all three tasks. The "risk neutral" task was always displayed first, with the other two following in a random ordering.

After each task, we asked the following:

1. How sure are you that you picked a movie that you will enjoy?
2. How sure are you that you made the best choice possible for this situation?
3. For this situation, how happy have you been with your MovieLens experience?
4. Did you treat movies with dice icons differently than movies without the icons?
5. If you answered "Yes" above, in what ways did you treat movies with dice icons differently?
6. For this situation, how valuable were the dice icons in helping you select a movie?

# 5 Results

New users logged in an average of 3.9 times during the experiment and rated an average of 215 movies each. Experienced users logged in an average of 6.6 times and rated an average of 19 movies during the experiment. The experienced users had already rated an average of 574 movies before joining the experiment.

## 5.1 Survey results

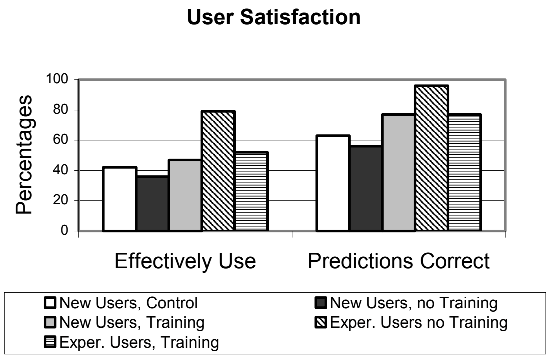
Most of the subjective questions were asked on a 5-point scale. In our results below, we display the results in bins for negative (1 or 2) or positive (4 or 5) answers. We dropped noncommittal responses.

Figure 4 shows the awareness users had of the confidence display. New users were less likely than experienced users to notice or understand the confidence display. Calling the icons out with training caused new users to notice them more frequently, but still not as often as experienced users, who noticed and understood the dice icons regardless of training. It is worth noting that experienced users also clicked the dice icons much more often. Untrained new users clicked on the dice icons an average of 0.3 times during the experiment, whereas those with training clicked 0.8 times. Untrained and trained experienced users clicked the dice 3.2 and 3.8 times on average, respectively.

Figures 5 and 6 show each group's satisfaction with MovieLens. New users with dice icons were roughly as satisfied as the control group, and those with training had higher satisfaction. Experienced users had much higher opinions of MovieLens than new users. This is not unexpected, since experienced users who did not like MovieLens probably stopped frequenting it a while ago.

The decline in user satisfaction for experienced users due to training was surprising. While training appears to be beneficial to new users, the introduction and training of a confidence display into the system had an adverse affect on experienced users. By explicitly calling out the possibility that not all recommendations are of the same accuracy, it is possible that the training planted a *seed of doubt* in these users' minds about all recommendations they have received or will receive from the system.

Table 2 shows the perceived value of the dice icons for the different experimental groups. Training tended to increase value for all users. Overall, experienced users also felt that the confidence display had more value than the new users. We think this is because experienced users were able to relate the confidence metric to their



percentages. New users who received training expressed more satisfaction with MovieLens than those who did not. Experienced users who received training, however, expressed less overall satisfaction.

prior experiences with MovieLens. As one such user said, "I liked the dice a lot; some of the predictions have always seemed a bit weird to me, [and] the dice made sense of it." Another one said, "I like the dice, helps me know why a particular movie *might* not be one I would like. So prefer to have that context when deciding what to see."

The new user control group seemed more enthusiastic about a hypothetical confidence display than the experimental group was about the actual feature. 47% (versus 6%) of the control group thought the dice icons would "Always be Valuable" and 37% (versus 20%) claimed they would be much more likely to check other sources of information about risky movies than non-risky movies.

| | New, no train | New, train | Exper, no train | Exper, train |
|---|---|---|---|---|
| Always Valuable | 0 | 6 | 18 | 26 |
| Never Valuable | 44 | 29 | 34 | 18 |

**Table 2:** Perceived value of the confidence display, in percentages. Both new and experienced users who received training tended to judge the confidence display as more valuable than their untrained counterparts.

It could just be that it's easier to believe that a feature is cool than to actually use it—the grass, as they say, is always greener. Or, it could be an indication that our specific metric or display decreased actual satisfaction below its potential.

Users had strong opinions about the addition of the confidence display. Over 50% of users provided comments, varying from, "I really liked the dice feature. It gives some interesting information about the database being used" to "If anything, the dice icons were just annoying, and seemed somewhat pointless to me." Many had insights and suggestions for improvement. One user thought, "it would be very informative, I think, if we could select any film in the database and see how many people had rated it," while another user, "would like to have the option to eliminate 'diced' entries entirely in my viewing."

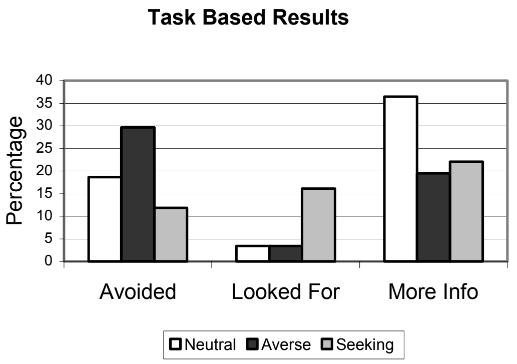


**Figure 7:** Task-based survey results. Users avoided movies with low confidence when performing a risk avoidance task, and looked for low-confidence movies when performing a risk-seeking task.

## 5.2 Task-Based Study Results

Confidence displays helped users distinguish and select movies for different tasks. As shown in Figure 7, users were more likely to avoid risky movies for the risk-averse scenario and more likely to seek them out for the risk-seeking scenario, compared with the risk-neutral scenario. They were also twice as likely to read extra information about a risky movie for the risk-neutral scenario, as compared with a non-risky movie. These results were consistent across both trained and untrained experienced users.

In general, users felt they would enjoy the movie that they chose in the risk neutral scenario more than the movie they chose for the other two scenarios. This makes sense, since the users were satisfying other constraints besides their personal enjoyment of the movie in the other scenarios. However, users remained consistent across the three scenarios when stating whether they felt they made the best choice possible for the given situation.

Since users were not told that the different scenarios equated to varying amounts of risk, the results are very encouraging about the ability of confidence displays to supplement predictions in helping users find items of interest, particularly for user tasks where the riskiness of the movie is an important factor.

Finally, the seeds of doubt carried over. As with overall satisfaction with MovieLens, experienced users who did not receive training were more satisfied with their movie choices than those who did. Untrained users thought they were more likely to have made the best choice possible for the situation (62% compared to 47%) and were happier with their selection (68% compared to 56%) than trained users.

# 6 Discussion

We summarize our primary results below:

- Adding a confidence display to a recommender system increases user satisfaction.
- Adding a confidence display to a recommender system alters users' behavior. For user tasks with varying amounts of risk, users were more likely to seek out or avoid low confidence recommendations as appropriate.
- New and experienced users do react differently to the addition of a confidence display. New users are not as likely to notice a confidence display, but will make use of it if they notice it. Experienced users all notice such a display, but already had opinions of the system, which affected their acceptance and use of the display.
- Training has a profound impact on user satisfaction in a recommender system. Providing training to new users increased user satisfaction over just adding the confidence display to the system. Providing training to experienced users increased their usage of the confidence system, but decreased their overall satisfaction with the recommender.

Training and timing of the confidence display both seem to have a substantial impact on user satisfaction and behavior. More importantly, their interaction causes the most profound impact. While the effect was noticeably positive for new users (as expected), the negative effect for experienced users deserves closer inspection. How could one page of training have that dramatic of an effect on the experienced users for happiness, effectiveness, and perceived correctness of predictions?

Swearingen and Sinha [2002] suggest that in order to develop trust in a recommender system, the functionality of the underlying algorithm needs to be transparent to the user. Without this transparency, users are left to their own imaginations to determine how a system operates. Many of the experienced users have spent a lot of time in MovieLens, with a mean of 574 movies rated (median of 435 ratings). These users have already developed mental models of how recommendations are generated for them.

By providing the confidence display, we explicitly pointed out that our collaborative filtering algorithms are not perfect. The users had no choice but to notice and accept this fact along with the confidence display and the transparency it offered, even if that clashed with their pre-existing models of the recommendation process. Further, the text we used for the training included the word "risky" twice and did not clearly explain that the dice icons simply represented movies with few ratings. Users who received the training may have had an emotional reaction to the notion of risk in the recommender system before understanding what the dice icons were meant to represent.

People treat computer systems and applications as social creatures [Reeves 1996], and when people invest time and energy in a system, they build trust relationships. People might view a recommender system as a trusted "individual" with whom they interact when they submit ratings and receive personalized recommendations. The training we provided for the confidence display didn't attempt to convey the recommender's "personality"; rather, we intervened in the user's relationship as experimenters. Further, we came in proclaiming that the computer had been giving "risky" recommendations, perhaps undermining the user's confidence. (How would you feel if your accountant suddenly had a supervisor watching to help you flag examples of "risky" accounting?)

In spite of all of this, Table 2 says the experienced users with training received the most value from the dice. These users may have had seeds of doubt planted in their worldviews, but they find real value in the confidence display. Perhaps a little doubt is justified when receiving low confidence recommendations. In our experiments, 49% of the recommendations presented to experienced users were for risky movies, compared to only 28% for the new users.[1] Thus, it is possible that the training offers real value and more transparency into the system—at the expense of reduced user confidence.

# 7 Conclusions and Future Work

[1] MovieLens does not show predictions for movies users have already rated. Experienced users had already rated many of the lower risk movies.

There is a tradeoff between having happy, but possibly naïve users who only extract part of the value from a recommender system and less happy but lucid users who are more successful at using the system. This tradeoff is a rich area to explore.

As a user grows in the system, it might be worthwhile to change the system's notion of risk for that user. Experienced users have already rated many items MovieLens can confidently predict for them. Thus, future predictions are likely to be less confident ones. Most of these users are currently happy, however, and might even be looking for these riskier recommendations. One user said, "I understood and appreciated the function of the dice icons, but they did not change my behavior in actually selecting movies to see. […] A majority of the movies I really like are not rated [by many people]."

The introduction of training in the user life cycle appears to be an important variable in users' overall satisfaction. As users interact with systems, they build mental models of how the system works and form an emotional bond with their software. Our results suggest that early training may lead to greater user happiness, since the users develop a more accurate model of the system.

Our study involved an elementary confidence computation, and simple associated interface displays. Since users saw these as valuable, investigating richer confidence computations and more sophisticated displays is worthwhile.

Adding a particular feature to an interface often seems an obviously good idea. As our study shows, the interaction between user experience, training, and the new features may be complex. In general, the problem of *how* to add features to an interface in a way that users can accept may turn out to be as difficult as deciding *what* to add—and is a problem that interface designers will often encounter as we strive to build interfaces which better serve users.

## 7.1 Acknowledgements

Tom Zielund performed related work for his M.S. project. We would like to thank Istvan Albert and Dan Cosley for their suggestions and comments. This work is supported by the NSF (DGE 95-54517, IIS 96-13960, IIS 97-34442, IIS 99-78717, and IIS 01-02229) and by Net Perceptions, Inc.

# References

Breese, J., Heckerman, D., and Kadie, C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In Proc. UAI 98, Madison, 1998, 43–52.

Canny, J. Collaborative Filtering with Privacy via Factor Analysis. In Proc. SIGIR 02, Tampere, Finland, 2002, 238-245.

Carroll, J. M. The Nurnberg Funnel: Designing Minimalist Instruction for Practical Computer Skills. The MIT Press, Cambridge, MA, 1990.

Carsewell, C., and Wickens, C. The Perceptual Interaction of Graphic Attributes: Configurality, Stimulus Homogeneity, and Object Interaction. In Perception & Psychophysics, 47: 157-168, 1990.

Herlocker, J., Konstan, J. A., Borchers, A., and Riedl, J. An Algorithmic Framework for Performing Collaborative Filtering. In Proc. SIGIR 99, Berkeley, 1999, 230–237.

Herlocker, J. Understanding and Improving Automated Collaborative Filtering Systems. PhD Thesis, University of Minnesota, 2000.

Herman, L., Ornstein, G., and Bahrick, H. Operator Decision Performance Using Probabilistic Displays of Object Location. In Engineering Psychology: Current Perspectives in Research, Howell, W., Goldstein, I. (editors). New York, Appleton-Century-Crofts, 1971, 69-74.

Mack, R. L., Lewis, C. H., and Carroll, J. M. Learning to Use a Word Processor: Problems and Prospects. ACM Transactions on Office Information Systems 1(3), 1983, 265-283.

McGrenere, J., Baecker R. M., and Booth, K. S. An Evaluation of a Multiple Interface Design Solution for Bloated Software. In Proc CHI 2002, Minneapolis, 2002, 164-170.

Reeves, B., and Nash, C. The Media Equation: How People Treat Computers, Television and New Media Like Real People and Places. New York: CSLI/Cambridge University Press, 1996.

Resnick, P., Iacovou, N., Sushak, M., Bergstrom, P., and Riedl, J. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In Proc. CSCW 94, Chapel Hill, 1994, 175–186.

Sarwar, B., Karypis, G, Konstan, J. A., and Riedl, J. Item-based Collaborative Filtering Recommendation Algorithms. In Proc. WWW 10, Hong Kong, 2001, 285–295.

Swearingen, K., and Sinha, R. Interaction Design for Recommender Systems. In Proc. DIS 2002, London, 2002.