Hypothetical Recommendation: A Study of Interactive Profile Manipulation Behavior for Recommender Systems

James Schaffer, Tobias Höllerer, John O'Donovan
Dept. of Computer Science,
University of California, Santa Barbara
{james_schaffer, holl, jod}@cs.ucsb.edu

Abstract

Explanation and dynamic feedback given to a user during the recommendation process can influence user experience. Despite this, many real-world recommender systems separate profile updates and feedback, obfuscating the relationship between them. This paper studies the effects of what we call hypothetical recommendations. These are recommendations generated by low-cost, exploratory profile manipulations, or "what-if" scenarios. In particular, we evaluate the effects of dynamic feedback from the recommender system on profile manipulations, the resulting recommendations and the user's overall experience. Results from a user experiment (N=129) suggest that (i) dynamic feedback improves the effectiveness of profile updates, (ii) when dynamic feedback is present, users can identify and remove items that contribute to poor recommendations, (iii) profile update tasks improve perceived accuracy of recommendations and trust in the recommender, regardless of actual recommendation accuracy.

Introduction

Recommender systems have evolved to help users get to the right information at the right time (Resnick et al. 1994; Sarwar et al. 1998). In recent years, a number of researchers and practitioners have argued that the user experience with recommendation systems is equally, if not more, important than accuracy of predictions made by the system (Herlocker et al. 2004). Research has shown that providing dynamic feedback to users during the recommendation process can have a positive impact on the overall user experience in terms of user satisfaction and trust in the recommendations and to accuracy of predictions (Bostandjiev, O'Donovan, and Höllerer 2012). In many real-world recommender systems, however, user profiles are not always up-to date when recommendations are generated, and users could potentially benefit from adding, removing, and re-rating items to reflect current preferences. While researchers have explored the effects of conversational recommender systems (McCarthy et al. 2005), these studies focus on a granular refinement of requirement specifications for individual product search during an ad-hoc session. In this paper, we focus on evaluating how the experience of the recommendation consumer is affected by using low-cost, exploratory profile manipulations (an addition, deletion, or re-rate) on a pre-existing profile to generate what we call "hypothetical" recommendations. These are scenarios that allow a user to update a stale profile by asking questions of the form "what if I added product x?", "what if I rated these 10 songs?" Specifically, this paper describes a study involving 129 participants, designed to answer the following three research questions:

1. How does dynamic feedback affect which type of profile updates users perform?
2. What is the effect of different types of profile updates on recommendation error?
3. What is the effect of dynamic feedback on *perceived* accuracy, satisfaction, and trust?

Previous work on profile elicitation for collaborative filtering systems has focused on passive (Rafter, Bradley, and Smyth 1999) and active (Boutilier, Zemel, and Marlin 2003) approaches. The experiments discussed in this paper can also be classed as a form of active profiling: the user is encouraged to add, delete, and re-rate items by assessing the feedback from the recommender while updating their preference profile. This research also considers the impact of interactive feedback for eliciting and encouraging profile manipulations from the user. Before we proceed with our discussion of the experiment itself, the following sections frame the experiment in the context of previous research on explanation and interaction aspects of recommender systems.

Engaging Users

The majority of research in recommender systems is focused on improving recommendation algorithms (e.g. (Koren, Bell, and Volinsky 2009)), without specific focus on user experience. This research builds on a number of related research efforts that deal with visualization, interaction and control of recommender systems. Earlier work by Herlocker (Herlocker, Konstan, and Riedl 2000) demonstrated that explanation interfaces for recommender systems can improve the user experience, increasing the trust that users place in the system and its predictions. Cosley et al. (Cosley et al. 2003) build on the explanation study to explore how explanations can change the opinions of a recommendation consumer, particularly in terms of rating behavior. They focus

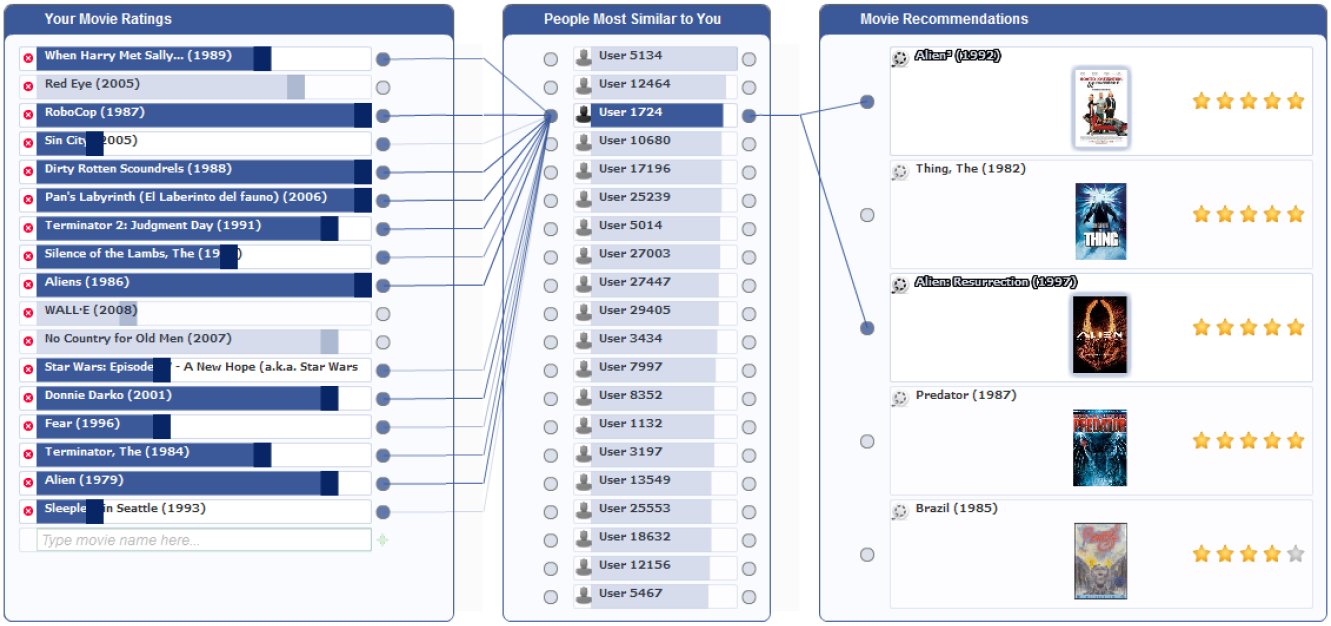

Figure 1: Screenshot of the interactive recommender system used in the experiment. From left to right the columns display: a user's profile items; top-k similar users; top-n recommendations. Adds, deletes or re-rates produce updated recommendations in real time. The dark blue lines appear when a user clicks a node and show provenance data for recommendations and the nearest neighbors who contributed to them. Clicking a movie recommendation on the right side of the page opens the movie information page on Rottentomatoes.com.

on consistency of re-rating behavior, impact of the rating scale and of dynamic feedback. Our experiment differs from Cosley's study (Cosley et al. 2003) in that feedback is not explicitly controlled to be high or low quality, placing the focus on the true impact of hypothetical profile manipulations on the overall user experience. Work by Swearingen and Singha (Sinha and Swearingen 2002) finds that users tend to have higher trust in recommender systems that predict items that they already know and like. They posit two important considerations for interaction design: what user needs are satisfied by interacting and what specific features of the system lead to satisfaction of those needs? In the context of our experiment, we believe that the user "need" is a desire to explore and probe the information space, and that a low-cost "hypothetical recommendation" feature provided by an interactive visualization tool can fulfill this user requirement.

Interactive Recommendation Systems

Recent work in this area focuses on visual interactive explanation and control mechanisms for recommender system algorithms. O'Donovan et al. (O'Donovan et al. 2008) describe an interactive visualization tool that supports genre-based manipulations of the k-nearest neighbors used in a collaborative filtering algorithm. They argue that "over-tweaking" can reduce the quality of recommendations if the interactive manipulations are not well balanced with the pre-existing user profile information. Bostandjiev et al. (Bostandjiev, O'Donovan, and Höllerer 2012) describe a visual interface to a hybrid recommender system that supports user guided transitions between social and semantic recommendation sources, and this system is leveraged by (Knijnenburg et al. 2012) in an experiment to study the effect of inspectability and control in social recommender systems. In particular, Knijnenburg et al. finds that both inspectability and control have a positive impact on user satisfaction and trust in the recommender system, but they do not evaluate this effect as profiles are manipulated over time . Verbert et al. (Verbert et al. 2013) further analyze the impact of information visualization techniques on user involvement in the recommendation process. Their evaluation of the Conference Navigator system (Verbert et al. 2013) shows that the effectiveness of recommendations and the probability of item selection increases when users are able to explore and interrelate entities.

Our work differs from previous approaches in that we attempt to determine the individual impact of each *type* of profile manipulation (add, delete, re-rate) on recommendation error, and how dynamic feedback affects the frequency of each type being performed and any difference in magnitude when reducing recommendation error.

Experimental Setup

In this study, the interactive recommender system shown in Figure 1 was presented to participants, and they were asked to add, delete or re-rate items in their profile. The system recommended movies based on the MovieLens 10M dataset, through two different configurations of the user interface.

Treatment	*First Phase*	*Second Phase*
1	Gathering	Manipulation (no dynamic feedback)
2	Gathering	Manipulation (w/ dynamic feedback)

Table 1: Breakdown of participant task and independent variables

Metric Name	*Explanation*
Manipulation	Participant's quantity of additions, deletions, and re-rates of profile items during the second phase of the task.
Rec. Error	Mean difference of ratings given by participants and ratings by the recommender.
Satisfaction	The participant's perceived satisfaction with the recommendations (1-100).
Trust	The participant's reported trust in the recommender (1-100).
Accuracy	The participant's perception of the accuracy of the recommender (1-100).

Table 2: Dependent variables in the study.

The first group received dynamic feedback (on-the-fly recommendations after each profile update) while the second group did not receive feedback. Pre-existing profile information was retrieved by participants through a web service of their choice (Netflix, IMDb, etc.) and we asked users to rate recommendations from the system based on this initial profile as a benchmark. Ratings were given on a 1-5 star scale. Following this, users updated their profiles using the interactive interface and received iterative feedback from the recommender based on a treatment (feedback or no feedback), and were subsequently asked to rate the post-manipulation set of recommendations from the system.

Design and Metrics

Participants in the dynamic feedback condition received recommendations generated by the system on the fly as they manipulated their profile in the second phase, while the remaining did not (see Table 1). By comparing ratings from the first phase against the second phase, and between treatments, we were able to examine how manipulation of profiles affected recommendation error, satisfaction, trust, and perceived recommendation accuracy in the presence and absence of dynamic feedback (Table 2). To use our earlier analogy, profile manipulations can be used to establish "what-if" scenarios at low cost to the user. Our aim is to assess how users go about this process, and what the resulting outcome is for their final recommendations and overall user experience.

The recommender system was deployed on Amazon Mechanical Turk (AMT) and data was collected from 129 AMT workers. Previous studies have established that the quality of data collected from AMT is comparable to what would be collected from supervised laboratory experiments, if studies are carefully set up, explained, and controlled (Buhrmester, Kwang, and Gosling 2011; Paolacci, Chandler, and Ipeirotis 2010). Previous studies of recommender systems have also sucessfully leveraged AMT as a subject pool (Bostandjiev, O'Donovan, and Höllerer 2012). We carefully follow recommended best practices in our AMT experimental design and procedures.

Generating Recommendations

Since the focus of this paper is on examining the profile manipulation behavior of users, and not specifically on the underlying recommendation algorithm, we chose a standard dataset (10m Movielens) and a standard collaborative filtering algorithm (Mobasher et al. 2007).

Algorithm

A collaborative filtering algorithm was chosen for this experiment because it lends itself well to visualization, but other algorithms should be interchangeable in the context of this experiment. Note that users have reported being able to easily understand visual representations of the algorithm (Bostandjiev, O'Donovan, and Höllerer 2012). The variant of collaborative filtering in this study applies Herlocker damping to increase recommendation quality.

User Interface

Our experiment uses a three column representation of collaborative filtering, similar to (Bostandjiev, O'Donovan, and Höllerer 2012). From left to right, a user sees his or her movie profile, then similar users in the collaborative filtering database, and finally a list of top movie recommendations. The underlying algorithm represents results from collaborative filtering as a directed graph, connecting the user's profile items to database users with at least one overlapping item and specifying edge strength as a similarity score. This score is shown as a light-blue gauge on the node for simplicity. Thus, if a user clicks on a movie he has rated, they can see which other similar users have rated it, and which recommendations are a result of those ratings. The recommendation column uses a star notation rather than a bar, provides visuals from the movie in the form of a teaser poster, and, when clicked, takes the user to RottenTomatoes.com to get more information about the movie.

Experimental Results

More than 300 users started the study, but many users were unable to complete the task properly due to the scarcity of valid "stale" profiles. Participant age ranged from 18 to 65, with an average of 31 and a median of 29. 53% of participants were male while 47% were female. Since we are interested in profile manipulation behavior, our experimental design did not enforce any minimum number of manipulations. After the initial profile collection phase, many users did not make enough updates to their profile, so could not be used in our analysis. Furthermore, some users indicated that their profile no longer required updates to accurately reflect their preferences, therefore implicitly indicating that the data was not truly 'stale' when the task was started. After

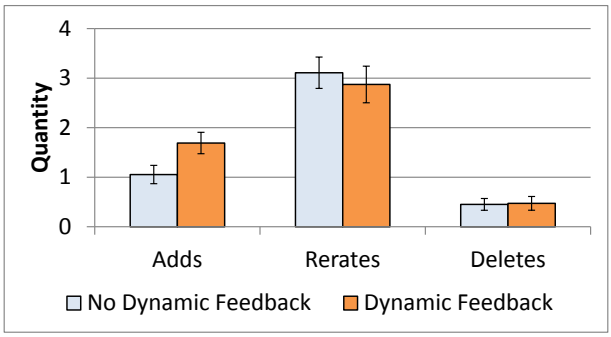

Figure 2: Frequency of each type of manipulation for each treatment. Error bars show one standard error below and above the mean. The most common action was re-rating an old item, and deletion of an old item was much more rare in comparison.

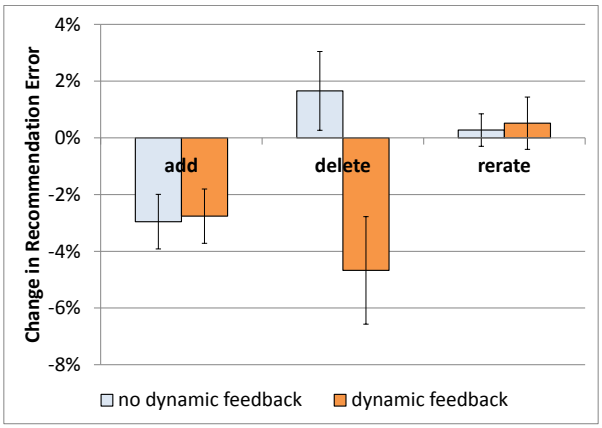

Figure 3: This graph shows the change in recommendation error that occurs from each type of manipulation in each treatment. These values were found by fitting a linear model to the manipulation patterns of participants that initially received poor recommendations. Error bars show one standard error below and above the mean. Adding new items was productive in both treatments, while deleting items was productive only in the dynamic feedback treatment.

removing these participants, data from 129 users (73 for the no feedback treatment, 55 for the feedback treatment) was analyzed. The average rating over initial recommendations for these users was 3.88 (out of 5) while the average rating for final recommendations was 3.93.

Effect of Dynamic Feedback on Profile Updates

After the user's profile was gathered, we allowed them to make an arbitrary of manipulations to update their profile and get hypothetical recommendations. A breakdown of the manipulation behavior, by treatment is shown in Figure 2. Re-rating a previously added item was the most common behavior in both conditions, followed by addition of a new item and deletion of an item respectively. Between both treatments, participants were 2.18x more likely to re-rate than add ($p < 0.01$), and 2.97x more likely to add than to delete ($p = 0.01$). Participants in the dynamic feedback treatment were also 1.6x more likely to add items than participants in the no feedback treatment with low presumption ($p = 0.108$).

Effect of Updates on Recommendation Error

Since we are interested in understanding the impact of each type of profile update on recommendation error, the error was measured with both the initial and final profiles so the two could be compared. Here, recommendation error is defined as mean absolute error ($MAE(p)$) for each participant p, or the difference between a participant's rating for an item and the predicted rating for that item:

$$MAE(p) = \frac{1}{n}\sum_{i=1}^{n}|p_i - r_i| \quad (1)$$

Where n is the total number of movies rated by participant p, p_i is the rating given by participant p to movie i, and r_i is the rating the system predicted participant p would give to movie i. Now we can define an error shift between the initial and final profiles of participant p by looking at the recommendations for each:

$$\delta error_p = MAE_{final}(p) - MAE_{initial}(p) \quad (2)$$

We realized one difficulty with our methodological approach is that users who initially received high quality recommendations are likely to exhibit different manipulation behavior from those with poor quality initial recommendation. Accordingly, we hypothesized that initial recommendation error and the resulting shift in error would have a significant interaction effect. We compensated for this by performing an analysis of the error shift based on the initial recommendation error. A linear regression showed that error shift was highly dependent on the initial recommendation error ($p < 0.01$). When the data is split on the average initial recommendation error (0.214), we find that users below the mean saw a 6.73% decrease in recommendation error after manipulation, while users above the mean saw a 1.14% increase ($p < 0.01$). In other words, users that had good initial recommendations could not do much to improve them, and in some cases manipulations caused increase in error, despite the fact that dynamic feedback was given during this process.

Given the above, we fit the following linear regression models to each treatment of users that saw initial recommendations with an error below the mean (no feedback: N=27, feedback: N=21):

$$\delta error(p) = adds(p) + rerates(p) + deletes(p) \quad (3)$$

Where $adds(p), rerates(p), deletes(p)$ return the quantity of those manipulations for participant p. The coefficients of the model indicate the impact of each type of manipulation had on recommendation accuracy for participants that had below average initial recommendations. We fit the

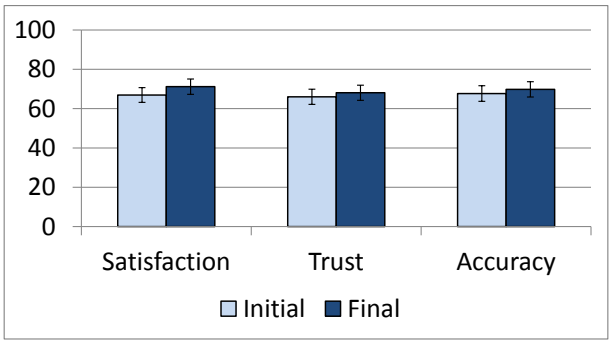

Figure 4: This graph shows participant responses to questions about satisfaction with recommendations, overall trust in the recommender, and perceived accuracy of recommendations for the no feedback treatment.

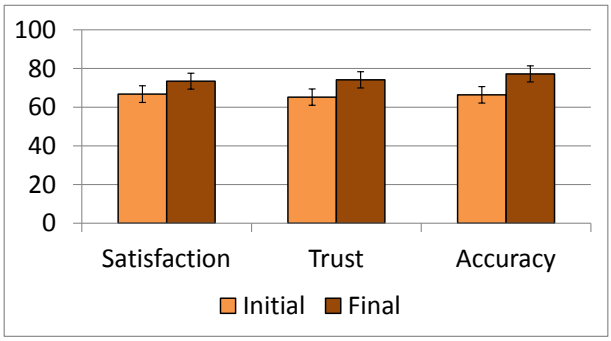

Figure 5: This graph shows participant responses to questions about satisfaction with recommendations, overall trust in the recommender, and perceived accuracy of recommendations for the dynamic feedback treatment.

regression model to both treatment groups and the resulting model coefficients are shown in Figure 3. Note that the model for the dynamic feedback group was accurately able to explain variability in the dataset ($p = 0.016, R^2 = 0.45$) vs. the model for the group without dynamic feedback ($p = 0.68, R^2 = 0.062$). The resulting models show that profile additions are the most effective manipulation for both treatments in terms of recommendation error, but deletes in the dynamic feedback group were the most effective manipulation overall. Deletes in the no-feedback treatment as well as re-rates in either treatment were either not effective or somewhat harmful to recommendation accuracy.

Effect of Dynamic Feedback on Perception

As stated before, perceptual metrics (overall satisfaction with recommendation, overall trust in the recommender, and perceived accuracy of recommendations) were taken after the final profile manipulation phase during the post-study test. Figure 4 and Figure 5 show a breakdown of the initial and final reports from each participant for these questions. We found that perceived trust and accuracy significantly increased for the dynamic feedback condition; this was verified by repeated-measures ANOVAs ($p = 0.0095$, and $p = 0.0158$ respectively). No significant change was found when dynamic feedback was not present. However, a mixed-measures ANOVA comparing the two treatments showed that the presence of dynamic feedback did not account for most of the change, as the before-and-after effect was more significant. Keep in mind this also applies to all participants in the dynamic feedback treatment, not just the ones that received poor initial recommendations, and that, overall, actual recommendation accuracy did not change for these participants.

Analysis and Discussion

Here we discuss the potential broader impacts of the results, limitations of the experiment, and plan for a follow-up study.

Re-rates are the most frequent type of profile manipulation, but have the least impact. To explain this effect, we posited that users did not change their rating very much from the initial to final profile. To verify this, we obtained the average difference between the original rating and any arbitrary re-rate, and found that most re-rates are within 1 point of the original rating on the 5 point scale provided. This supports the idea that, for movies at least, user tastes do not change very much over time. Visual recommenders in similar settings might consider the option of foregoing the functionality to re-rate, or perhaps put more emphasis on adding items or removing them altogether.

Dynamic feedback and visual explanation let users identify sources of bad recommendation and remove them. In our dynamic feedback treatment, delete actions improved recommendation accuracy by more than 4% on average for each individual delete performed. The most likely explanation is that users identified bad recommendations and were able to use the interface to determine and remove the item causing the correlation.

Users overvalue their profile updates. When users updated their profiles, they perceived that the overall accuracy of the recommendations and trust in the system increased significantly, even though the actual recommendation error stayed the same on average. While our mixed-measures ANOVA showed no significant effect of dynamic feedback on perception, participants in the no feedback condition reported on average that there was more or less no difference in accuracy and trust from the initial and final profiles. Thus, we still recommend that if a service requires a user to perform a profile update task (such as the first time a user accesses the system after a long period), dynamic feedback should be utilized.

Limitations and Future Work

The authors note several points where this study could have been improved. First, there was difficulty in acquiring truly stale profiles from users and obtaining quality manipulations. As stated before, many users that started the study copied their profile into our system and then skipped the manipulation phase of the study. We considered enforcing that participants make some threshold number of manipulations, but any such enforcement would prevent our measurement of true profile manipulation behavior. A second limi-

tation was that we only used a single recommendation strategy (collaborative filtering) in this experiment. It is not clear whether the findings about the manipulation types would apply to other recommendation algorithms or to different data domains. Additionally, the choice of MovieLens 10M was for ease of implementation, and we note that many participants requested more up-to-date movie recommendations. Finally, our rating system could have been improved by considering list-based satisfaction, since disjoint ratings do not fully capture user satisfaction.

A larger follow up study is planned to gather more profile manipulation data from participants. The study will also examine the effects of profile manipulation and hypothetical recommendations across different recommendation algorithms (e.g. content-based, collaborative and matrix factorization approaches) and will be applied to two data domains (movie and career data) to study the portability of our research results. Additionally, our follow-up study will use a more complete and up-to-date dataset for better applicability of results.

Conclusion

This paper described an experiment (N=129) to evaluate the impact of different types of low-cost, exploratory manipulations on a preference profile for collaborative filtering recommender systems. The experiment tested one condition in which dynamic feedback on profile manipulations was provided, and one with no feedback. Our data supports the following claims: (i) the presence of dynamic feedback elicits marginally more profile additions, (ii) the addition of new items to a profile reduces recommendation error, and when dynamic feedback is present, deletes become significantly more effective, and (iii) profile update tasks improve perceived accuracy and trust, regardless of any change in actual recommendation error.

References

Bostandjiev, S.; O'Donovan, J.; and Höllerer, T. 2012. Tasteweights: a visual interactive hybrid recommender system. In Cunningham, P.; Hurley, N. J.; Guy, I.; and Anand, S. S., eds., *RecSys*, 35–42. ACM.

Boutilier, C.; Zemel, R. S.; and Marlin, B. 2003. Active collaborative filtering. In *In Proceedings of the Nineteenth Annual Conference on Uncertainty in Artificial Intelligence*, 98–106.

Buhrmester, M.; Kwang, T.; and Gosling, S. D. 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6(1):3–5.

Cosley, D.; Lam, S. K.; Albert, I.; Konstan, J. A.; and Riedl, J. 2003. Is seeing believing?: How recommender system interfaces affect users' opinions. In *Proceedings of the Conference on Human Factors in Computing Systems*, 585–592. ACM Press.

Herlocker, J. L.; Konstan, J. A.; Terveen, L. G.; John; and Riedl, T. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems* 22:5–53.

Herlocker, J. L.; Konstan, J. A.; and Riedl, J. 2000. Explaining collaborative filtering recommendations. In *Proceedings of ACM CSCW'00 Conference on Computer-Supported Cooperative Work*, 241–250.

Knijnenburg, B. P.; Bostandjiev, S.; O'Donovan, J.; and Kobsa, A. 2012. Inspectability and control in social recommenders. In Cunningham, P.; Hurley, N. J.; Guy, I.; and Anand, S. S., eds., *RecSys*, 43–50. ACM.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer* 42(8):30–37.

McCarthy, K.; Reilly, J.; McGinty, L.; and Smyth, B. 2005. Experiments in dynamic critiquing. In *IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces*, 175–182. New York, NY, USA: ACM Press.

Mobasher, B.; Burke, R.; Bhaumik, R.; and Williams, C. 2007. Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Trans. Inter. Tech.* 7(4):23.

O'Donovan, J.; Smyth, B.; Gretarsson, B.; Bostandjiev, S.; and Höllerer, T. 2008. Peerchooser: visual interactive recommendation. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, 1085–1088. New York, NY, USA: ACM.

Paolacci, G.; Chandler, J.; and Ipeirotis, P. G. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision Making* 5:411–419.

Rafter, R.; Bradley, K.; and Smyth, B. 1999. Passive profiling and collaborative recommendation. In *Proceedings of the 10th Irish Conference on Artificial Intelligence and Cognitive Science, Cork, Ireland*. Artificial Intelligence Association of Ireland (AAAI Press).

Resnick, P.; Iacovou, N.; Suchak, M.; Bergstrom, P.; and Riedl, J. 1994. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM CSCW'94 Conference on Computer-Supported Cooperative Work*, 175–186.

Sarwar, B. M.; Konstan, J. A.; Borchers, A.; Herlocker, J.; Miller, B.; and Riedl, J. 1998. Using filtering agents to improve prediction quality in the grouplens research collaborative filtering system. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, 345–354. ACM.

Sinha, R., and Swearingen, K. 2002. The role of transparency in recommender systems. In *CHI '02 extended abstracts on Human factors in computing systems*, 830–831. ACM Press.

Verbert, K.; Parra, D.; Brusilovsky, P.; and Duval, E. 2013. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, 351–362. New York, NY, USA: ACM.