Impact of Item Consumption on Assessment of Recommendations in User Studies

Benedikt Loepp, Tim Donkers, Timm Kleemann, Jürgen Ziegler
University of Duisburg-Essen
Duisburg, Germany
{firstname.lastname}@uni-due.de

ABSTRACT

In user studies of recommender systems, participants typically cannot consume the recommended items. Still, they are asked to assess recommendation quality and other aspects related to user experience by means of questionnaires. Without having listened to recommended songs or watched suggested movies, however, this might be an error-prone task, possibly limiting validity of results obtained in these studies. In this paper, we investigate the effect of actually consuming the recommended items. We present two user studies conducted in different domains showing that in some cases, differences in the assessment of recommendations and in questionnaire results occur. Apparently, it is not always possible to adequately measure user experience without allowing users to consume items. On the other hand, depending on domain and provided information, participants sometimes seem to approximate the actual value of recommendations reasonably well.

CCS CONCEPTS

• **Information systems → Recommender systems**;

KEYWORDS

Recommender Systems; Experimentation; User Studies

1 INTRODUCTION AND RELATED WORK

Measuring user experience becomes increasingly important in *Recommender Systems* (RS) research [9, 12, 13, 16], with user studies playing an important role for evaluating system quality, especially in academia [10]. In these studies, participants typically use a system and are subsequently asked to fill in a questionnaire [7, 10]. Based on items such as "I liked the products recommended by the system" [11], researchers then draw inferences about e.g. perceived recommendation quality. Recommended products are usually represented by textual descriptions, pictures and metadata. Only in rare cases, it is possible to actually consume them [e.g. 19]. Consequently, participants often have to judge recommendations based on limited knowledge. In real-world scenarios, e.g. when people want to rate a product or write a hotel review, it is in contrast often required to have bought a product or visited a hotel before providing an opinion. This led us to the following questions: *What is the impact of item consumption on the assessment of recommendations in RS user studies? Are there domain-specific differences that determine whether users can adequately assess the value of recommendations without experiencing the items?* We conducted two user studies in different domains, music and movies, to investigate pre- and post-consumption assessments of recommendation quality and aspects related to user experience of RS. As usual, recommendations were presented with descriptive data, but we also enabled participants to consume items, i.e. listen to songs or watch movies.

Most related to this paper is the research on explanations [2, 19–23]. Among others, it has been found that users over-/underestimate recommended items depending on type and quality of explanations [21]. Yet, in the studies conducted, it was typically not possible for participants to consume products. In some cases, this was at least approximated: In [2], Amazon detail pages of recommended books could be read. In the experiments described in [22], watching movies was only possible in one case, while reviews were shown otherwise. The authors compared before- and after-ratings, but put their focus on the effectiveness of explanations provided in addition to a very limited result presentation. Another exception is the study in [19], where participants could listen to song recommendations that were explained. Other works have shown, for instance, that the point in time preferences were elicited plays an important role as users provide lower ratings the longer ago an item was experienced [3]. In [1], the authors investigated anchoring effects in rating behavior: No differences were found between predicted ratings being presented as anchors before or after watching a TV show. As participants were only asked for their opinion after consumption, the actual influence of experiencing items remained unclear. Differences in user behavior and consistency of ratings can, however, have a considerable effect on RS performance [17]. Also, the presentation of recommended items is long known for its impact [4]. In [6, 14], movie recommendations were accompanied by trailer videos. Yet, the authors did not analyze whether availability of such options affects user experience, and thus the resulting questionnaire responses. In summary, examining possible differences in the assessment of recommendations and of related aspects between before and after item consumption in RS user studies is therefore an important open research topic.

2 USER STUDY 1: SONGS

We hypothesized that actually listening to recommended songs makes a difference in how users assess subjective system aspects [10, 11] such as perceived recommendation quality, and aspects related to user experience such as satisfaction with the chosen

RecSys '18, October 2–7, 2018, Vancouver, BC, Canada

item. We assumed there would be intra-individual differences in pre- and post-consumption assessment of recommendations, but also differences between users depending on whether they had consumed recommended items prior to the assessment.

2.1 Method

We set up condition S1 with questionnaires before (Pre) and after (Post) consumption, and S2, with a questionnaire only afterwards (Post). We conducted a controlled experiment with 40 participants (22 female), average age of 26.00 ($SD = 8.69$), a small majority of them students (65 %). We assigned them to conditions in counter-balanced order in a between-subject design ($N_{S1} = 21$, $N_{S2} = 19$). Participants reported liking music a lot ($M = 4.05$, $SD = 1.04$). Yet, 28 % did not know any of the recommended songs, the rest only a few ($M = 1.42$, $SD = 1.30$). For recommending and playing songs, we implemented a web application using the Spotify API.

Procedure: First, participants in both conditions had to select 3 out of 110 Spotify genres. Next, in S1, they were presented with a list of 5 recommendations (generated using the API with selected genres as seed data). Song titles, artists, album titles and covers were displayed. Participants were required to rate their satisfaction with each recommendation and fill in the questionnaire. Then, the recommendation list was shown in both conditions (again in S1). Participants were asked to listen to each song for at least 30 sec with the possibility to stop, pause and forward. Finally, they (again in S1) had to rate the recommendations and fill in the questionnaire.

Questionnaire: For composing the questionnaire, we relied on established RS evaluation instruments. We used constructs from [11, 15] that have been shown to operationalize system aspects and user experience reasonably well with a limited number of questions. We also generated items ourselves to ask whether participants were in doubt when selecting recommendations and which criteria they found most influential. In S1-Pre, we also asked how likely they would change their ratings when they could listen to songs, and in S1-Post, which reasons they had to change them (open-ended question). All items were assessed on a 1–5 Likert-scale.

2.2 Results and Discussion

We fitted linear mixed-effect models for each dependent variable, measured by one or more questionnaire items, with condition (S1, S2) and point in time (Pre, Post) as a fixed factor, specified point in time as a repeated measurement, and conducted custom hypothesis tests for fixed effect parameters. Tab. 1 shows the comparison of S1-Pre with S1-Post (i.e. within-subject) and S1-Pre with S2-Post (i.e. between-subject). We omit results for S1-Post vs. S2-Post as we found significance only for choice satisfaction (0.57 better in S2-Post, $SE = 0.23$, $p = .018$), confirming that participants had the same knowledge in both conditions after listening.

Within-subject effects: Performing custom hypothesis tests in case of significant interaction terms confirmed, among others, that participants gave higher ratings to recommendations after listening to the recommended songs (cf. Fig. 1). Questionnaire results are in line: Prior to consumption, perceived rec. quality shows a significant correlation with mean rec. rating ($r = .603$, $p = .004$). Other constructs, e.g. overall satisfaction ($r = .575$, $p = .006$), correlate as well, validating our results. After consumption, we found similar correlations. Moreover, the difference of mean rec. rating between the two assessments is larger, the lower perceived rec. quality in S1-Pre ($r = -.709$, $p < .000$). Listening had apparently more influence when recommendations were initially perceived less appropriate, while participants saw few need to change ratings otherwise. Also, ratings are normally distributed with a variance of 0.77 in S1-Pre, while the distribution is bounded with less variance of 0.33 in S1-Post (cf. Fig. 1, left). This suggests that participants before consumption had difficulties to form a strong opinion [22], but were more certain afterwards.

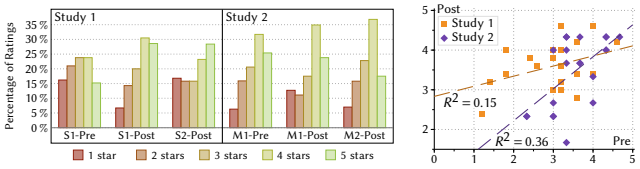

Figure 1: Distribution of ratings for recommendations (left) and scatter plot of mean recommendation ratings (right).

Choice and overall satisfaction are higher in S1-Post (Tab. 1). Since participants were already asked to settle for a preferred item in S1-Pre, choice difficulty was lower when they were confronted with the same list again in S1-Post. Mean rec. rating tends to correlate between the two points in time ($r = .390$, $p = .080$, see Fig. 1, right), which is consistent with prior work on re-rating of items [4, 8]. Overall, Fig. 1 underlines that scores are on average higher after consumption. We found similar correlations for choice ($r = .510$, $p = .018$) and overall ($r = .600$, $p = .004$) satisfaction, i.e. questionnaire results correlate as well. Also, listening to songs had a significant effect on the perceived sufficiency of information provided for recommendations (Tab. 1). The difference between S1-Pre and -Post is higher, the fewer items are known ($r = -.492$, $p = .023$), i.e. in typical RS where novelty is pursued, consumption seems especially important to support users. While the artist was the most influential criterion in S1-Pre, followed by song title and album cover, listening was considered most influential in S1-Post, underlining the need for consuming items from a subjective perspective. Qualitative comments support this: Participants reported that "artist or album title are not meaningful, while it is important to like how a song sounds" and that "listening allowed imagining how well a song fits the own taste". Others wrote: "I had bad expectations when I read the artist's name, but was positively surprised when I heard the song" or "I knew one song, but could only remember and rate it after listening". Overall, participants were more satisfied when they found information sufficient in S1-Pre ($r = .745$, $p < .000$). When sufficiency was low, they had more doubts ($r = .463$, $p = .034$). The more doubts, the more they reported they would change their ratings due to listening ($r = -.682$, $p = .001$). This effect was generally very high ($M = 4.52$, $SD = 0.87$).

Between-subject effects: Hypothesis tests comparing S1-Pre and S2-Post show slightly fewer differences (Tab. 1), likely because there were no sequential persuasion effects as in S1-Post, where participants had already seen recommended items. We found only one significant correlation of the (small) number of items known before our study (as expected, with the difference of information sufficiency between S1-Pre/-Post, see above). Since item familiarity

is similar in S1 and S2, it thus does not seem to be a confounding factor. The influence of anchoring effects on the differences found within subjects also appears negligible, which is supported by absence of differences between S1- and S2-Post.

Participants rated system effectiveness, choice and overall satisfaction higher when they could listen to songs prior to filling in the questionnaire (Tab. 1). Choice difficulty tended to be lowered as well, which in the within-subject case, however, was likely due to design. Subjective system aspects, e.g. perceived rec. quality, and mean rec. rating were also not significant. Apparently, participants were able to judge the recommendation set without listening, but had difficulties to correctly assess the aforementioned aspects related to user experience. Overall, this shows that the typical design of RS studies may contribute to an inaccurate picture compared to when users can experience items.

Questionnaire results are again aligned with mean rec. rating. In S2-Post, there is a correlation with perceived rec. quality ($r = .796$, $p < .000$), which, in turn, and in line with earlier work [10], correlates with overall satisfaction ($r = .546$, $p = .016$). Experiencing the songs also led to higher perceived information sufficiency and fewer doubts (Tab. 1). In addition, variance in ratings is lower than in S1-Pre as well (0.77 vs. 0.53). Accordingly, while 81 % in S1-Pre stated it would be very useful to listen to songs or at least extracts, the most influential factor in S2-Post was listening, followed by artist and album cover. Participants listened only a little longer in S2 ($M = 70.4$ sec, $SD = 41.7$) than in S1 ($M = 53.6$ sec, $SD = 20.7$), without significance or relevant correlations. Again, they were overall more satisfied, the higher information sufficiency ($r = .626$, $p = .004$), and the fewer songs were known ($r = -.565$, $p = .012$): Recommending known items not only conflicts with the RS goal of adding novelty, but, moreover, consuming them seems to decrease user satisfaction. Interestingly, in S1-Pre ($r = .462$, $p = .035$), and still S1-Post ($r = .427$, $p = .054$), this correlation is reversed, likely because assessing a recommendation set is easier in case related items are known, which might induce bias in typical RS studies.

3 USER STUDY 2: MOVIES

3.1 Method

The second study was designed similar to study 1, with conditions M1 and M2. We again recruited 40 participants (30 female) with average age of 21.78 ($SD = 3.77$), a large majority of them students (92 %). We assigned them to conditions as in study 1 ($N_{M1} = 21$, $N_{M2} = 19$). They reported liking movies ($M = 3.63$, $SD = 0.98$). As item data, we used 13 short movies available at YouTube recommended in an online article of the German newspaper *Zeit* (http://bit.ly/zeit-movies). Only 2 participants knew one of the movies before.

Procedure: First, participants had to provide demographics and select one category out of "Horror, Mystery & Thriller", "Comedy & Romance" or "Drama". Then, they were presented with a list of 3 pseudo movie recommendations (chosen randomly from the selected category). We displayed movie titles, genres, posters, metadata on director and cast, and (subjective) description texts by the article's author. Next, in M1, participants were required to rate their satisfaction with each recommendation and fill in the questionnaire. Afterwards, in M1 and M2, they had to select a movie they would like to watch. Only in M1, they had to answer questions regarding this choice (t_1). Then, in both conditions, they had to watch this movie (entirely, without pausing/forwarding). After watching, they had to rate (re-rate in M1) their satisfaction with this recommendation and answer corresponding questions (t_2). Next, they had to watch and assess the two remaining movies. Eventually, participants again had to choose one movie (independent of their previous choice) and answer questions regarding this choice (t_3).

Questionnaire: The questionnaire was similar to study 1. Due to slightly different design, questions regarding the chosen item were now asked separately, and thus three times in M1 (at t_1, t_2 and t_3).

3.2 Results and Discussion

We fitted mixed models as in study 1. Tab. 1 shows comparisons within (M1-Pre vs. M1-Post) and between (M1-Pre vs. M2-Post).

Within-subject effects: Only few interaction terms are significant, with custom hypothesis tests showing no differences between M1-Pre and -Post. Still, mean ratings for individual recommendations are in line with questionnaire results: Before consumption, perceived rec. quality shows a significant correlation ($r = .515$, $p = .017$). Overall satisfaction highly correlates with mean rec. rating as well ($r = 634$, $p = .002$). Afterwards, correlations were even stronger.

As in study 1, mean rec. rating also correlates well between points in time ($r = .600$, $p = .004$, see Fig. 1, right). At the same time, Fig. 1 aligns with questionnaire results, i.e. there are no differences in ratings (only a slight decrease with more 1-star ratings in M1-Post, but also more 4-star ratings). In comparison to study 1 with smaller correlations but significant differences with respect to ratings and questionnaire responses, the richer information seemed to make it easier for participants to form a strong opinion, i.e. deviate from the scale midpoint. Thus, while the distribution became bounded in study 1 only after consumption, this was already the case in M1-Pre (similar to [22]). Overall, this underlines participants indeed can assess recommendations consistently—depending on domain and opportunities to approximate their value, e.g. by means of subjective descriptions as taken from the newspaper. Questionnaire results support these findings, e.g. perceived rec. quality correlates as well between assessments ($r = .660$, $p = .001$). Yet, in contrast to study 1, the score in M1-Pre does not seem to affect the difference found with respect to mean rec. rating ($r = .157$, $p = .496$).

The most influential information in M1-Pre was the newspaper description, followed by poster, genre and title. Some participants reported that "information was too basic, directors not helping (all unknown) and casts not allowing to conclude about movie quality" so that they "relied entirely on description and genre". Although quantitative results do not differ, they chose a different item when this was possible at t_3 in 62 % of all cases. Indeed, this might be influenced by participants who assumed that they would have to watch the movie again, and wanted to circumvent this. Still, selection was altered more often with lower perceived rec. quality prior to consumption ($r = -.470$, $p = .031$), and lower satisfaction with initial choices after watching the respective movie (at t_2, $r = -.548$, $p = .010$). The significant interaction for choice satisfaction (Tab. 1) is related to this and the final assessment at t_3: After seeing the two other movies (and possibly changing the selection), the assessment was estimated to be 0.57 higher ($SE = 0.20$, $p = .011$). This might be attributable to participants being more convinced of

Table 1: Results of our mixed models for both user studies for interaction of condition (S1, S2/M1, M2) and point in time (Pre, Post). Positive differences indicate better results after consumption (*Choice Diff.*, *Effort* and *Doubts* are reversed accordingly).

Study 1	Interaction	S1-Pre vs. S1-Post			S1-Pre vs. S2-Post		
	Sig.	Est. Diff.	Std. Err.	Sig.	Est. Diff.	Std. Err.	Sig.
Perceived Rec. Quality [11]	.390	0.38	0.28	.183	0.15	0.29	.611
Mean Recommendation Rating	**.009***	**0.59**	**0.18**	**.004***	0.30	0.24	.226
Choice Satisfaction [11]	**.000***	**0.71**	**0.21**	**.003***	**1.29**	**0.28**	**.000***
Choice Difficulty [11]	**.001***	**1.14**	**0.29**	**.001***	0.55	0.38	.156
Effort [11]	.415	0.21	0.16	.196	0.10	0.23	.664
Effectiveness [11]	**.000***	**0.81**	**0.19**	**.000***	**1.08**	**0.33**	**.002***
Diversity [11]	.056	-0.38	0.26	.151	0.42	0.31	.184
Novelty [15]	.288	-0.19	0.13	.144	0.11	0.30	.731
Information Sufficiency [15]	**.000***	**1.48**	**0.38**	**.000***	**1.67**	**0.38**	**.000***
Transparency [15]	.104	0.48	0.22	.051	0.61	0.38	.113
Confidence and Trust [15]	**.017***	**0.54**	**0.20**	**.014***	**0.64**	**0.26**	**.020***
Doubts	**.000***	**2.19**	**0.33**	**.000***	**1.71**	**0.38**	**.000***
Overall Satisfaction [15]	**.005***	**0.62**	**0.20**	**.005***	**0.89**	**0.31**	**.007***

Study 2	Interaction	M1-Pre vs. M1-Post			M1-Pre vs. M2-Post		
	Sig.	Est. Diff.	Std. Err.	Sig.	Est. Diff.	Std. Err.	Sig.
Perceived Rec. Quality [11]	.467	-0.14	0.17	.411	-0.27	0.27	.328
Mean Recommendation Rating	.771	-0.08	0.14	.578	-0.11	0.21	.574
Choice Satisfaction [11]	**.020***	-0.19	0.25	.450	0.03	0.35	.937
Choice Difficulty [11]	.968	0.05	0.31	.877	-0.05	0.37	.905
Effort [11]	**.012***	-0.07	0.08	.383	**-0.47**	**0.15**	**.003***
Effectiveness [11]	.479	-0.14	0.22	.520	-0.41	0.34	.229
Diversity [11]	.117	0.24	0.19	.224	-0.37	0.34	.288
Novelty [15]	.218	0.14	0.09	.106	0.14	0.20	.472
Information Sufficiency [15]	**.041***	-0.33	0.23	.149	-0.37	0.32	.250
Transparency [15]	.763	-0.14	0.21	.499	-0.16	0.36	.658
Confidence and Trust [15]	.787	0.04	0.16	.826	-0.18	0.28	.527
Doubts	.680	-0.14	0.27	.605	-0.29	0.35	.407
Overall Satisfaction [15]	.442	-0.14	0.22	.525	-0.36	0.30	.235

their choice after experiencing all movies, but also to those who updated their selection to the movie they liked most after watching. The larger difference between these assessments in case a different movie was chosen ($r = .567$, $p = .007$) corroborates this assumption. We found similar results in M2.

Participants reported, similar to study 1, that "only after watching, it became clear which movies were of most interest, while descriptions were not sufficient to decide" and that "from the movie initially chosen, more was expected after reading its description". However, this time, numerous participants stated that "the watching experience met the expectations raised from the provided information", "ratings remained constant as descriptions allowed to get a pretty good impression" and "summaries helped to quickly grasp what to expect, making eager to watch the movies".

Between-subject effects: When comparing M1-Pre with M2-Post, only a single hypothesis test yields significance (Tab. 1). However, we already expected effort to be higher as participants in M1-Pre immediately rated recommendations and filled in the questionnaire, while in M2-Post, they had to watch all three movies ex ante. Questionnaire results are again in line with mean rec. rating: As in all aforementioned cases, we found a significant correlation in M2-Post with perceived rec. quality ($r = .617$, $p = .005$), which in turn correlates with overall satisfaction ($r = .789$, $p < .000$, cf. [10]).

In summary, there seems to be no considerable effect of introducing the possibility to consume recommended items prior to assessing the respective recommendations. This is clearly in contrast to study 1, and suggests that domain as well as type and amount of provided information determine whether an adequate picture of RS user experience can be obtained. Probably, it is naturally more easy in the movie than in the music domain to comprehend why certain items are recommended, even without consumption. Moreover, the information shown was richer (newspaper texts were more informative than song metadata) and more subjective (including the author's opinion and going beyond typical plot descriptions). Comments in the within-comparison support that "descriptions corresponded well to movies" and were "written subjective and emotionally". One participant explicitly stated that "the description revealed so much, there was no reason to change the movie's rating after watching it". Overall, it seems participants were able to accurately estimate whether they will like recommended items.

4 CONCLUSIONS AND OUTLOOK

In conclusion, it seems necessary to take questionnaire results of RS user studies with a grain of salt. Participants in some cases cannot adequately assess all aspects of a RS, especially those related to user experience, without consuming recommended items. Although we measured some concepts only by means of one or two suggested key questions [11], possibly limiting validity [10], this generalization appears reasonable. For instance, we found that participants in the music domain tend to underrate songs and are less satisfied when choosing from a list of recommendations that only contains descriptive information, instead of being able to listen to songs. The latter had a positive influence on satisfaction, leading to significantly higher scores for related questionnaire items. Since research on choice satisfaction has shown people tend to overestimate the impact of past events [25], and in certain circumstances become less satisfied after a few weeks [24], investigating stability of these results will be necessary. Subjective system aspects such as perceived rec. quality were rated equal independent of consumption. For movies, this seems true in more aspects, especially if high-quality textual descriptions are available, which is more likely the case for movies than for music with its abstract emotional content. Indeed, the fact that songs were more often known might have introduced bias. Yet, while short movies were nearly completely unknown, the number of known songs was also relatively low and yielded almost no correlations. Still, item familiarity and its impact on assessing recommendations should be studied in more detail. Also, the general influence of extensively using questionnaires needs to be considered more, as thinking consciously about decisions was found not always beneficial [5].

Overall, quantitative results indicate that it highly depends on domain as well as type and amount of information provided alongside recommendations whether the actual experience can sufficiently be substituted. Qualitative comments in both studies reflect this. Thus, we suggest to avoid comparisons across different settings and to pay attention in user experiments when requiring participants to rate recommendations without consumption. Likewise, item ratings provided in real-world RS that do not prevent rating e.g. movies or recipes before having seen or cooked them should be combined carefully with ratings elicited afterwards. Only when presenting adequate information, participants' responses may be reliable: In this case, item consumption may not be needed as they seem to form a mental model in which they also take their own preferences into account, allowing to judge recommendations the same way as if items really had been experienced. Otherwise, study results appear to provide at least a lower bound, which is particularly relieving since there are domains where it is not feasible to let participants consume recommended items, e.g. due to time or cost constraints (full movies, books, hotels). Nevertheless, we will further investigate the influence of provided information in dependence of product type (search/experience [18]) and domain.

REFERENCES

[1] Gediminas Adomavicius, Jesse C. Bockstedt, Shawn P. Curley, and Jingjing Zhang. 2013. Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects. *Information Systems Research* 24, 4 (2013), 956–975.

[2] Mustafa Bilgic and Raymond J. Mooney. 2005. Explaining Recommendations: Satisfaction vs. Promotion. In *Proceedings of the Beyond Personalization Workshop.*

[3] Dirk Bollen, Mark P. Graus, and Martijn C. Willemsen. 2012. Remembering the Stars? Effect of Time on Preference Retrieval from Memory. In *RecSys '12: Proceedings of the 6th ACM Conference on Recommender Systems.* ACM, 217–220.

[4] Dan Cosley, Shyong K. Lam, Istvan Albert, Joseph A. Konstan, and John Riedl. 2003. Is Seeing Believing? How Recommender Interfaces Affect Users' Opinions. In *CHI '03: Proceedings of the 21st ACM Conference on Human Factors in Computing Systems.* ACM, 585–592.

[5] Ap Dijksterhuis and Zeger van Olden. 2006. On the Benefits of Thinking Unconsciously: Unconscious Thought can Increase Post-Choice Satisfaction. *Journal of Experimental Social Psychology* 42, 5 (2006), 627–631.

[6] Mark P. Graus and Martijn C. Willemsen. 2016. Can Trailers Help to Alleviate Popularity Bias in Choice-Based Preference Elicitation?. In *IntRS '16: Proceedings of the 3rd Joint Workshop on Interfaces and Human Decision Making for Recommender Systems.* 22–27.

[7] Asela Gunawardana and Guy Shani. 2015. *Recommender Systems Handbook.* Springer US, Chapter Evaluating Recommender Systems, 265–308.

[8] Will Hill, Larry Stead, Mark Rosenstein, and George Furnas. 1995. Recommending and Evaluating Choices in a Virtual Community of Use. In *CHI '95: Proceedings of the 13th ACM Conference on Human Factors in Computing Systems.* ACM, 194–201.

[9] Michael Jugovac and Dietmar Jannach. 2017. Interacting with Recommenders – Overview and Research Directions. *ACM Transactions on Interactive Intelligent Systems* 7, 3 (2017), 10:1–10:46.

[10] Bart P. Knijnenburg and Martijn C. Willemsen. 2015. *Recommender Systems Handbook.* Springer US, Chapter Evaluating Recommender Systems with User Experiments, 309–352.

[11] Bart P. Knijnenburg, Martijn C. Willemsen, and Alfred Kobsa. 2011. A Pragmatic Procedure to Support the User-Centric Evaluation of Recommender Systems. In *RecSys '11: Proceedings of the 5th ACM Conference on Recommender Systems.* ACM, 321–324.

[12] Joseph A. Konstan and John Riedl. 2012. Recommender Systems: From Algorithms to User Experience. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 101–123.

[13] Benedikt Loepp, Catalin-Mihai Barbu, and Jürgen Ziegler. 2016. Interactive Recommending: Framework, State of Research and Future Challenges. In *EnCHIReS '16: Proceedings of the 1st Workshop on Engineering Computer-Human Interaction in Recommender Systems.* 3–13.

[14] Theodora Nanou, George Lekakos, and Konstantinos Fouskas. 2010. The Effects of Recommendations' Presentation on Persuasion and Satisfaction in a Movie Recommender System. *Multimedia Systems* 16, 4-5 (2010), 219–230.

[15] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-Centric Evaluation Framework for Recommender Systems. In *RecSys '11: Proceedings of the 5th ACM Conference on Recommender Systems.* ACM, 157–164.

[16] Pearl Pu, Li Chen, and Rong Hu. 2012. Evaluating Recommender Systems from the User's Perspective: Survey of the State of the Art. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 317–355.

[17] Alan Said and Alejandro Bellogín. 2018. Coherence and Inconsistencies in Rating Behavior: Estimating the Magic Barrier of Recommender Systems. *User Modeling and User-Adapted Interaction* (2018).

[18] Sylvain Senecal and Jacques Nantel. 2004. The Influence of Online Product Recommendations on Consumers' Online Choices. *Journal of Retailing* 80, 2 (2004), 159–169.

[19] Amit Sharma and Dan Cosley. 2013. Do Social Explanations Work? Studying and Modeling the Effects of Social Explanations in Recommender Systems. In *WWW '13: Proceedings of the 22nd International Conference on World Wide Web.* ACM, 1133–1144.

[20] Nava Tintarev and Judith Masthoff. 2008. The Effectiveness of Personalized Movie Explanations: An Experiment Using Commercial Meta-Data. In *AH '08: Proceedings of the 5th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems.* Springer, 204–213.

[21] Nava Tintarev and Judith Masthoff. 2008. Over- and Underestimation in Different Product Domains. In *Proceedings of the ECAI Workshop on Recommender Systems.* 14–19.

[22] Nava Tintarev and Judith Masthoff. 2012. Evaluating the Effectiveness of Explanations for Recommender Systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 399–439.

[23] Nava Tintarev and Judith Masthoff. 2015. *Recommender Systems Handbook.* Springer US, Chapter Explaining Recommendations: Design and Evaluation, 353–382.

[24] Timothy D. Wilson, Douglas J. Lisle, Jonathan W. Schooler, Sara D. Hodges, Kristen J. Klaaren, and Suzanne J. LaFleur. 1993. Introspecting about Reasons can Reduce Post-Choice Satisfaction. *Personality and Social Psychology Bulletin* 19, 3 (1993), 331–339.

[25] Timothy D. Wilson, Jay Meyers, and Daniel T. Gilbert. 2003. "How Happy was I, Anyway?" A Retrospective Impact Bias. *Social Cognition* 21, 6 (2003), 421–446.