Explaining Content-Based Recommendations

Bruce Ferwerda b.ferwerda@student.tue.nl
Kevin Swelsen k.j.m.swelsen@student.tue.nl
Emily Yang e.yang@student.tue.nl

Abstract

When confronted with too many choices, our satisfaction decreases (Iyengar & Lepper, 2000). Recommender systems may help increase satisfaction by making certain items more salient than others. Content-based recommendations are based on previously rated items of similar content and therefore can be more accurate. Recently, a framework for the user experience of recommender systems suggests that subjective system aspects influence user experience in addition to objective system aspects (Knijnenburg, Willemsen, & Kobsa, 2011b). In addition, previous studies (Herlocker, Konstan, & Riedl, 2000; Pu & Chen, 2007) suggest that explanations will increase trust and therefore increase user satisfaction on recommender systems. A study using a content-based TV program recommender system was used to evaluate whether content-based recommendations and explanations increased user satisfaction and other correlated behaviors. The study did not find significant results for content-based recommendations increasing satisfaction and other correlated behaviors. As for explanations, significant results were found partially. Although the condition with explanations did not outperform the one without, an increasing positive effect can be noticed for users with higher levels of activity.

1.0 Introduction

With so many choices these days, we have a hard time making decisions. Iyengar and Lepper (2000) have demonstrated that indeed, when confronted with too many choices, our satisfaction decreases. As number of options increase, people become less happy with their choice as they feel that they probably did not pick the best option.

These days, many websites help users make choices by using recommender systems to provide a smaller set of options that are likely to be of interest to that user. Recommender systems have been around for a while. E-commerce companies such as Amazon.com recommends products to their customers based on what others have purchased. Recommender systems are being applied to more than just online experiences; these days they are also implemented in set-top boxes giving users recommendations on what TV-program to watch. By implementing recommender systems, websites can hope to increase *choice satisfaction*.

1.1 ProgramGenie: TV program recommender

A company in the Netherlands has produced a TV program recommender system. For confidentiality reasons, the name *ProgramGenie* is used for this system for this report. ProgramGenie is a TV-guide, showing which programs are scheduled at what channel and when. It is built for set-top boxes, iPhone and for the World Wide Web. ProgramGenie is unlike other TV program because it is personal and content-based. First, programs are recommended by demographic information, then ProgramGenie gives recommendations based on previously rated programs and shows, which should increase user satisfaction as they are more tailored for the user's preference. Explanations for these recommendations may also increase user satisfaction.

1.2 Content-based recommendations

A content-based recommender system recommends products according to how a user rated associated features of other products (Burke, 2002; Degemmis et al., 2004, as cited by Ochi, Rao, Takayama, & Nass, 2009). For example, if a user has rated Bermuda highly and New York City poorly on a travel destination website, a content-based recommender system could use characteristics of these places such as climate and population density to recommend the Bahamas and not Chicago.

1.3 Demographic-based recommendations

Content-based recommender systems need input before it can work. ProgramGenie requires its users to rate programs before it infer content-based recommendations. Therefore, it utilizes user demographics to cold-start. For example, a 25-year-old female might initially be given *Gossip Girl,* a popular TV show for her demographic, as a recommended program before she rates anything. However, once she rates programs such as *Desperate Housewives* and *The Vampire Diaries* poorly, *Gossip Girl* will no longer be recommended for her.

By including a content-based algorithm, ProgramGenie should increase user satisfaction, as demographic-based recommendations are not as accurate as content-based recommendations

1.4. Recommender Systems: A Black Box

Regardless of algorithm method, one problem with recommender systems is that users may not understand what exactly is going on. This can be referred as a black box model as users see recommendations, but do not know how the system produced those recommendations. For example, such a model can result in a perceived lack of control with the system and therefore result in lower satisfaction with the system. When users cannot understand how a system works, those who are less trustful tend to want more control of the system (Vries, 2004, as cited by Knijnenburg et al., 2011b). Currently, ProgramGenie is a black box as users see a rating but do not see how the system arrived at that conclusion. This can be seen in figure 1.

Figure 1. Details of a program from condition <u>without</u> explanation

Just as people trust recommendations of friends who know their tastes and not that of a stranger, people may trust systems that they know their tastes and can easily explain how it came up with such recommendations. Sinha and Swearingen (2002) found that when users like and feel more confident about recommendations that they perceive as more *transparent*. Muramatsu and Pratt (2001, as cited by Sinha & Swearingen, 2002) found similar results when they studied user mental models for query transformations by search engines during the retrieval process. By making the process transparent to users, search performance was improved. With transparent systems, users have better mental models of systems. Thus, making it easier for users to adjust their input to get desired outputs if necessary.

With content-based systems like ProgramGenie, it may be beneficial for users to understand that an item is recommended because it has similar characteristics as a previously rated item. This may enable the user to better trust the system and better control the outcome of the recommendations.

1.5 Evaluation Framework for User Experience of Recommender System

Previously, researchers have focused on creating better algorithms to help produce better recommendations, which in turn help produce better user experience. However, Knijnenburg, Willemsen, Gantner, Soncu, and Newell (2012) argue that subjective system aspects (SSA), in addition to objective system aspects (OSA), of a recommender system may influence user experience (EXP) as summarized in figure 1. The inclusion of OSA, such as explanations, for recommendations may influence how people perceive the system (SSA) and therefore how their overall experience and subsequent behaviors. In addition, personal characteristics (PC), such as their propensity to trust systems or their persistence in finding the best option, may influence overall user experience (Knijnenburg, Reijmer, & Willemsen, 2011a).

Including explanations may increase interaction usability, perceived quality, appeal, and trust, which may in turn increase positive experience with the system.

In addition, content-based recommendations should produce better user satisfaction compared to demographic-based recommendations because it is personalized and therefore more accurate.

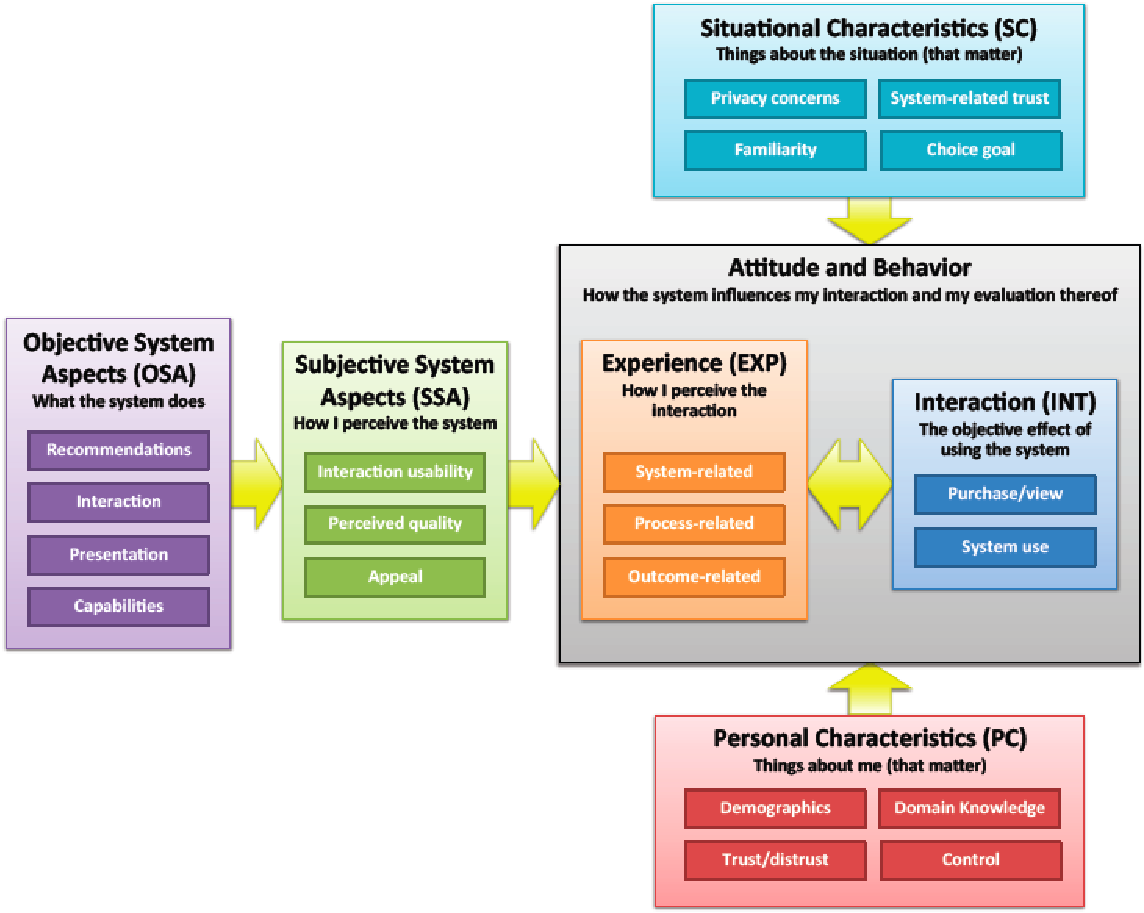

Figure 2. Recommender System Framework (Knijnenburg et al., 2012)

1.6 Explaining Recommendations: Towards a White Box Model

One way to increase transparency is to explain how a recommender system inferred the connection from the user to the product. Herlocker et al. (2000) did this with a collaborative filtering recommender system and found that users valued explanations.

Explanations must also be carefully constructed as their effectiveness at reducing cognitive effort can increase trust (Pu & Chen, 2006). An ineffective explanation may lead to an ineffective perception of the system and therefore lose users' trust. Pu and Chen (2007) have found that when recommendations are organized in a clearer and efficient manner, users trust in the systems increased as their perception of the system's competence and ability to explain its recommendation increased.

1.7 Explaining ProgramGenie Recommendations

Herlocker et al. (2000) tested several explanation methods for recommender systems. The most convincing design was a histogram. Therefore, we used histograms to help explain the ProgramGenie program ratings. For example, Tim, an 18-year-old fan of director Quentin Tarantino would see that *Pulp Fiction* is rated "8". The explanation will show that this is because the movie has Tarantino as its director (content) and that Tim has previously rated positively towards other Tarantino films; therefore, a positive Tarantino rating contributed to the rating of "8". An example of ProgramGenie with explanations can be seen in figure 3.

Figure 3. Details of a program from condition <u>with</u> explanation

In addition, some of the characteristics (content) of the program had a negative correlation, so the histogram was able to show low ratings for certain characteristics. For example, not only would Tim see that *Dora the Explorer* (a children's TV show) is rated 1 for him, he would also see why. Upon closer look on the explanation, he would see that the characteristic "children" was on the low end of the rating scale. This would make sense as he has previously rated other children's show with a "1".

2.0 Hypotheses

We will evaluate the ProgramGenie recommender system and the use of explanations with regard to user experience and correlated behavioral measures. Factors for evaluations were gathered by a previous study by Knijnenburg, Willemsen, Soncu, Newell, and Gantner (2012). Overall usability attitudes were measured using the Questionnaire For User Interaction Satisfaction (QUIS). User persistence was based from the Maximizing Satisfaction Scale (Knijnenburg et al., 2011a). The following list is a summary of each factor:

- User Interaction Satisfaction (QUIS) – how satisfied users are with the system (EXP)
- Choice Satisfaction – how satisfied users were with their chosen item (EXP)
- Perceived Usefulness/System Effectiveness – how effective and useful the system was with their goal (EXP)
- Perceived System (Recommendation) Quality – how well the system works with regards to recommendation fit (SSA)
- Trust – how well users believe that recommendations are reliable (SSA)
- Understandability – how easy it was to understand the system (SSA)
- User persistence – the tendency for a user to search thoroughly for the best option (maximize) as opposed to quickly settling for a good enough option (satisfice) (PC)

Using the framework by Knijnenburg et al. (2012), user experience (EXP) is measured by User Interaction Satisfaction, Choice Satisfaction, and Perceived System Quality. Trust and Understandability are subjective system aspects (SSA) and user persistence is a personal characteristic (PC). In this study, Trust is more about perceived reliability and accuracy and therefore not a situational characteristic as it has been previously categorized in other studies.

Based on Iyengar and Leppar's work (2000), we believe that by making a few programs salient, ProgramGenie's recommendations will improve user satisfaction. Content-based recommendations should be more accurate as demographic-based predictions are rough generalizations and with more accuracy, the system should be perceived as having better quality. The following hypothesis is illustrated in figure 4.

H1: <u>Content-based recommendations</u> should enhance Perceived System Quality (SSA) which increases user experience (i.e. User Interaction Satisfaction, Choice Satisfaction, Usefulness - EXP) and interaction (Number of Ratings - INT) with the system.

We expect that when explanations are available, users will be more satisfied because the system is perceived as more transparent as Sinha and Swearingen (2002) have found in their studies. With more transparency and understanding, users should then be able to trust the system more and have better control of their outcomes. With trust and understanding as SSA, users should be able perceive that the system is more useful, that the system has better quality, and that they feel more satisfied with their choices as with the system as a whole. The path for the following hypothesis can also be seen in figure 4.

H2: Explanations (OSA) should increase trust and understandability, which should increase the user experience (i.e. user interaction satisfaction, choice satisfaction, usefulness - EXP) and interaction (number of ratings - INT) with the system

Additionally, we will measure user persistence because user characteristics can determine how satisfied users are with different recommender interfaces (Knijnenburg et al., 2011a). User persistence is how persistent a user will maximize (search thoroughly for the best option) as opposed to satisficing (quickly settle for a good enough option). Maximizers try hard to find the best option possible with the least amount of trade-offs. Explanations make it easier for maximizers to find the best options. The path for the following hypothesis can also be seen in figure 4.

H3: User experience (EXP) and interaction measures will be greater for maximizers when explanations are present.

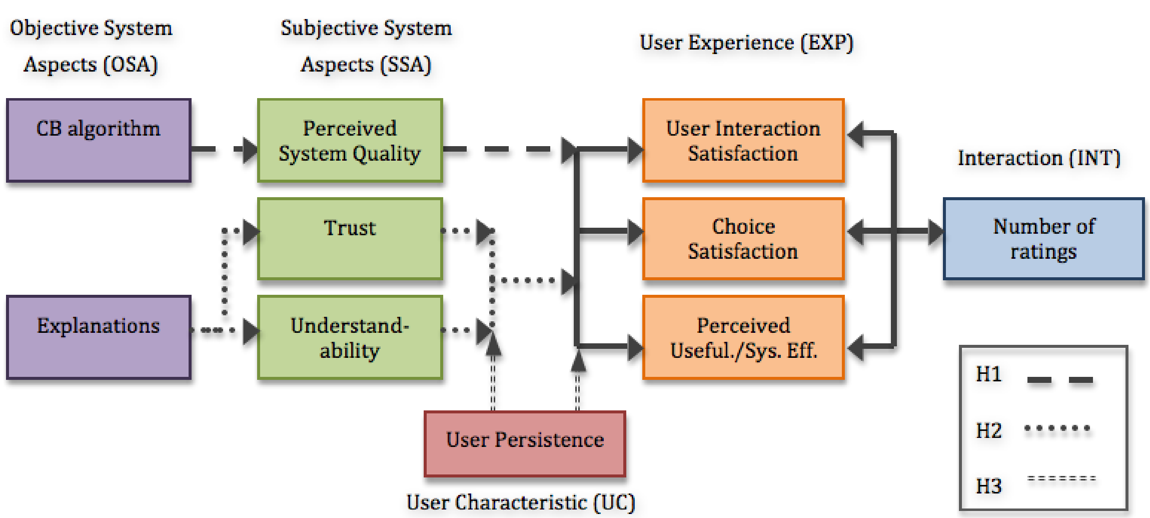

Figure 4. Model of hypotheses using the framework from Knijnenburg et al., 2012.

3.0 Methods

Participants were asked to use the ProgramGenie website for two weeks. No specific instructions were given; participants were allowed to do whatever they wanted to do on the website. During the experiment, participants were asked to fill out two questionnaires; one after the first week of participation, the other after the second week. As an incentive, €25,- was raffled for every tenth participant who participated the full two weeks.

3.1 Design

In the experiment, OSA was manipulated. Therefore, the following 2x2 design was defined: content-based recommendations (on/off) x explanations (on/off). When content-based recommendations were off, recommendations were given based on user demographics.

All conditions were in Dutch. A total of four copies of the website were made for the experiment, one version for each condition. Except for whether users were shown explanations, each condition had identical visual appearances and perceived functionality. Actual system functionality differed between conditions; although perceived functionality was the same, each condition was paired with different set of functionality and presence of explanation.

1. Content-based off + Explanations off
    - (Baseline) The system continues to use demographic-based recommendations. Participants were able to rate TV-programs, but their ratings did not have any effect on future recommendations.
2. Content-based off + Explanations on
    - This was same as condition 1, but with explanations turned on.
3. Content-based on + Explanations off
    - Participants received demographic-based recommendations at first. After rating, the recommendations adjusted according to the user's preference. Ratings will have an effect on future recommendations.
4. Content-based on + Explanations on
    - This was the same as condition 3, but with explanations turned on.

Evaluations were conducted by having participants answer questionnaires, in Dutch, measuring: subjective user experience and usability attitudes. Furthermore, for each participant, the system logged their behavior on the website.

3.2. Participants

Participants were recruited through a company specializing in marketing surveys in the Netherlands. A total of 312 participants registered for participation. Of these 312 participants, 105 participants were willing to participate by registering on the website. Only 97 participants used the website and filled out the questionnaire for a single week. The mean age was 47.27, ranging from 17 to 84. Of the participants, 64 were women and 41 men. Participants reported that they, on average, watched 34 hours of TV each week. Only 11 participants participated for both weeks: showed activity on the website and filled out the questionnaire for both weeks. Due to the lack of participants that participated two weeks, only participants of a single week are taken into account (N = 86). The 11 participants that participated both weeks are excluded.

3.3 Procedure

The initial 312 registered participants were equally divided between the four conditions. An e-mail was sent to all participants with instructions and a link to register on the specific ProgramGenie website corresponding to the condition. No further instructions were provided except for how to register.

After a few days, an email was sent to all participants to remind those who did not register yet and to remind registered users to make everyday use of the system.

No specific scenarios or tasks were given for the website. Users were able to browse through the website the way they preferred. After the first and second week, users were sent an e-mail with instructions to fill out a questionnaire (Appendix A).

4.0 Results

A regular factor analysis was conducted. An initial analysis was run to obtain eigenvalues for each component in the data and to check for cross and low factor loadings. QUIS is a standardized scale and therefore excluded from the factor analysis. Nine components had eigenvalues over Kaiser's criterion of 1 and in combination explained 79.02% of the variance. The scree plot was slightly ambiguous and showed inflexions. Of the original 38 items, 19 items were deleted with high cross and low factor loadings. 19 items remained for the final analysis, which resulted in seven components with eigenvalues over Kaiser's criterion of 1 and in combination explained 80.15% of the variance.

A principal component analysis (PCA) was conducted on 19 items with oblique rotation (direct oblimin). The Kaiser–Meyer–Olkin (KMO) measure verified the sampling adequacy for the analysis, KMO = .798. Bartlett's test of sphericity χ^2 (325) = 1525.475, $p <$ .001, indicated that correlations between items were sufficiently large for PCA. Of the nine components in the initial analysis, five components remained after rotation (Appendix B). The items that cluster on the same components suggest the following factors:

1. Trust
2. Perceived system quality
3. Perceived usefulness/System effectiveness
4. Choice satisfaction
5. Understandability

To investigate the effects of the different conditions on the factors, separate ANOVA's have been conducted with the extracted factors on algorithm (on/off) and explanations (on/off). As mentioned before only 11 participants that participated the whole two weeks. Therefore, it is decided to only include participants in the analyses that participated a single week (N = 86). Due to the extreme lack of activity, the dichotomous scale was used to divide people into a low and high activity group. Separation of groups is based on their times of login (>2) and their amount of ratings (>15) indicating high (N = 27) or low activity (N = 59). The activity has been added as a third factor in the ANOVA's. Furthermore, we have tested the effect of the amount of ratings as a covariate. Due to the exponential distribution of the ratings a log (rating+.5) data transformation has been used to correct for this. Participants of the whole two weeks are used in the analysis. No significant effects have been found on QUIS and Perceived usefulness/System effectiveness. Significant results across other factors are reported below.

4.1 Trust

Results indicate (figure 5) that the presence of explanations has a marginal significant negative effect on the amount of trust people have in the system, $F(1, 86) = 3.378$, $p = .07$. When explanations are not present, the mean and standard error are $M = .204$ and $SE = .160$. When explanations are present, the mean and standard error are $M = -.248$ and $SE = .187$. In other words, trust in the system decreases when explanations are shown. Results show no significant difference between algorithm on and off; trust in both conditions are statistically the same.

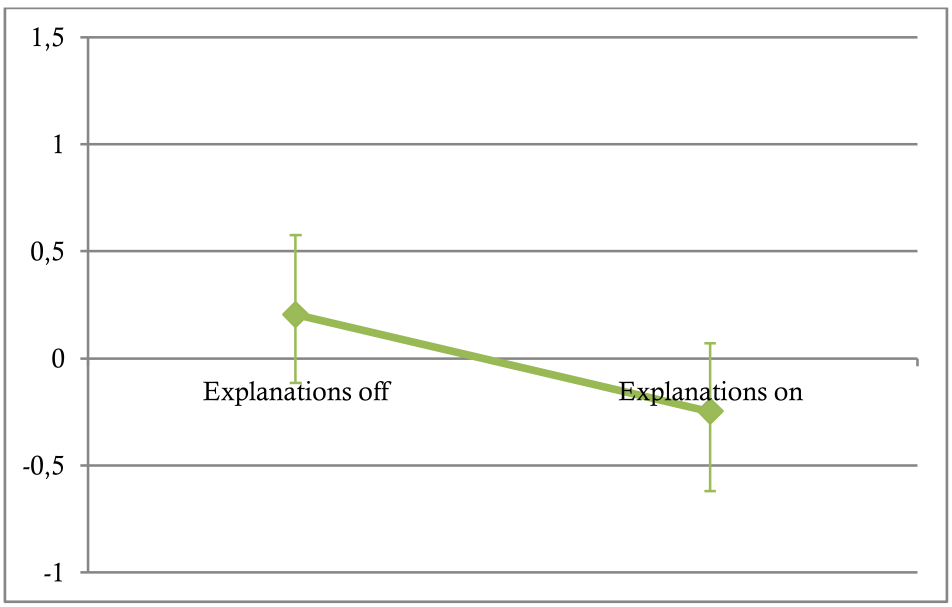

Figure5. Trust: Explanation off vs. Explanation off, 1 SE

However, a near significant interaction effect was found between explanation, algorithm, and activity, $F(1, 86) = 2.212$, $p = .12$. There was still a negative effect on trust when explanations and the content-based algorithm was present among low activity users. However, the mean and standard error of the interaction effect indicate an increase in trust when people showed more activity. High activity users (figure 7) show the most reduced trust when explanations are on with algorithm off. This might indicate that people noticed that recommendations with algorithm off do not deliver good quality recommendations when explanations are shown. This suggests that trust improves when people are using the system more frequently.

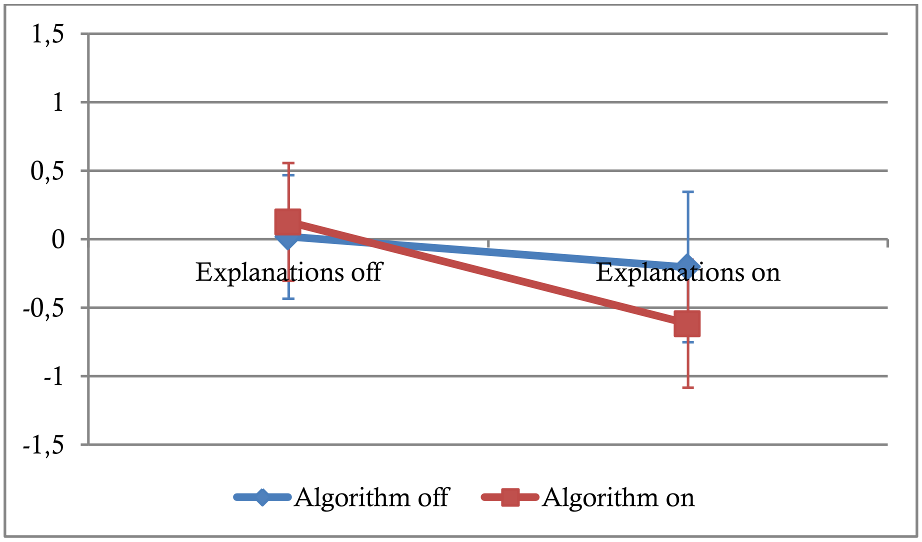

Figure6. Trust with low activity users, 1 SE

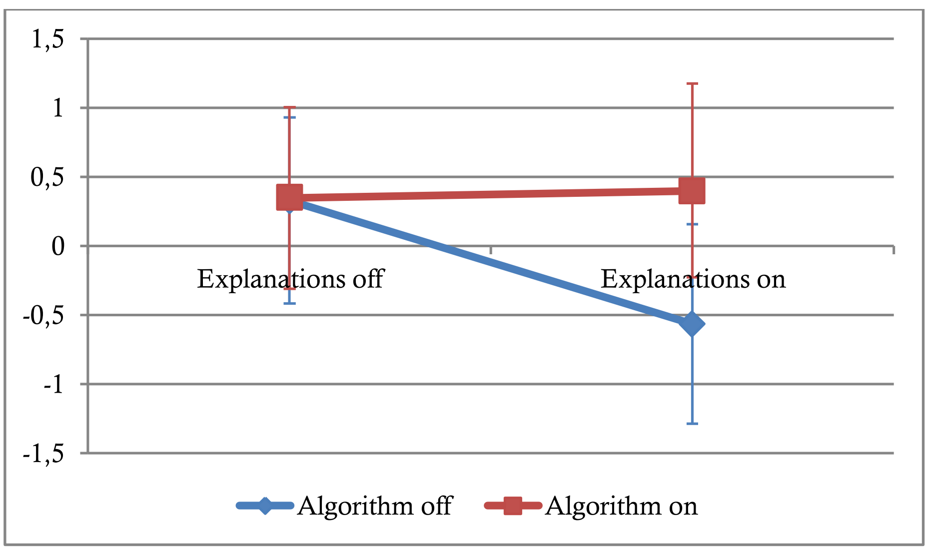

Figure7. Trust with high activity users, 1 SE

4.2 Perceived System Quality

Although no main and no interaction effect of explanations on/off and algorithm on/off were found, results indicate that there is a significant difference between people with low and high activity on perceived quality of the system (figure 8), $F(1, 86) = 4.510$, $p = .026$. People with higher activity generally perceived the system of higher quality.

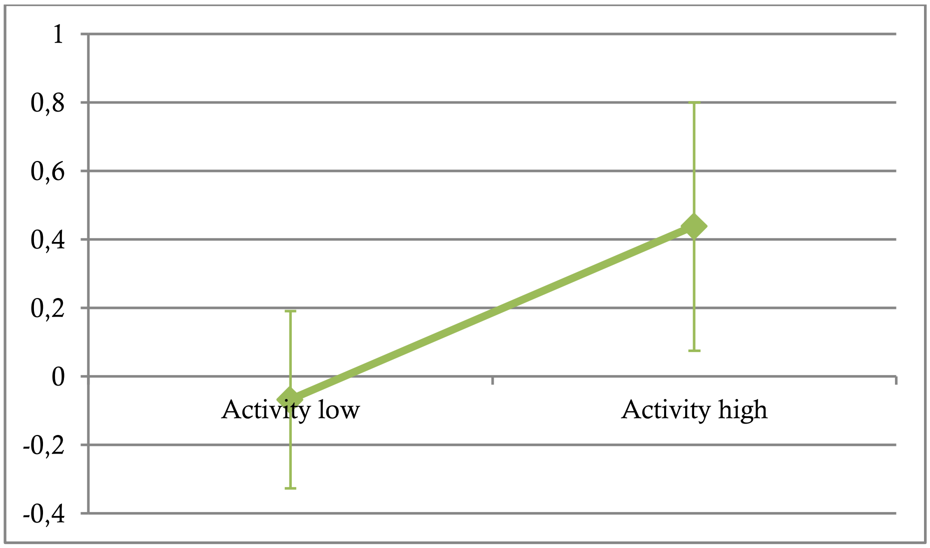

Figure 8. Perceived System Quality: Activity low vs. Activity high, 1 SE

4.3 Choice Satisfaction

A near significant interaction effect has been found between explanations (on/off) and algorithm (on/off), $F(1, 86) = 2.501$, $p = .107$. Results show that there is an increase in choice satisfaction when explanations are shown. However, a working algorithm seems to not contribute to this. People show a higher choice satisfaction when explanations are shown in combination with algorithm off (figure 9).

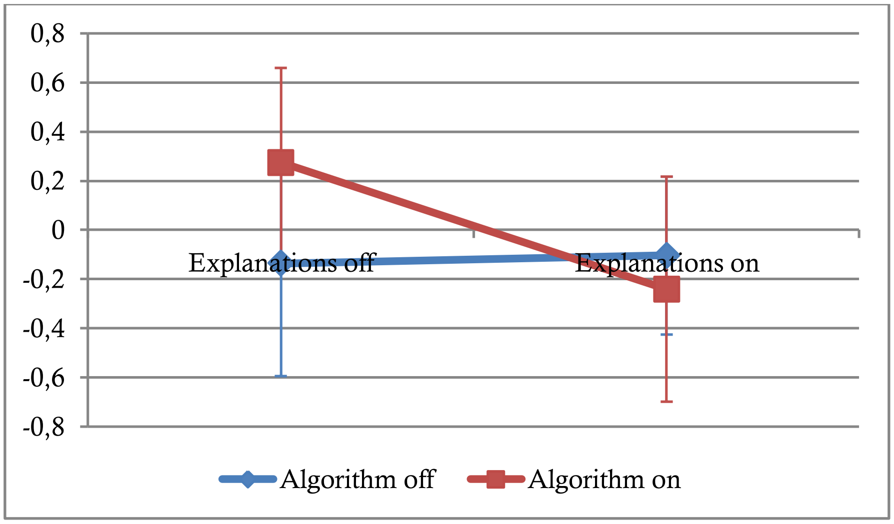

Figure9. Choice Satisfaction, 1 SE

However, when taking activity level into account, a significant main effect on choice satisfaction was found, $F(1, 86) = 5.459$, $p = .022$. This indicates that there is a significant difference between people with low and people with high activity. Choice satisfaction increases for people with higher activity (figure 10). When accounting for activity in the ANOVA, the marginal significant interaction between algorithm and explanation disappears. No further effects of explanations and algorithm are found.

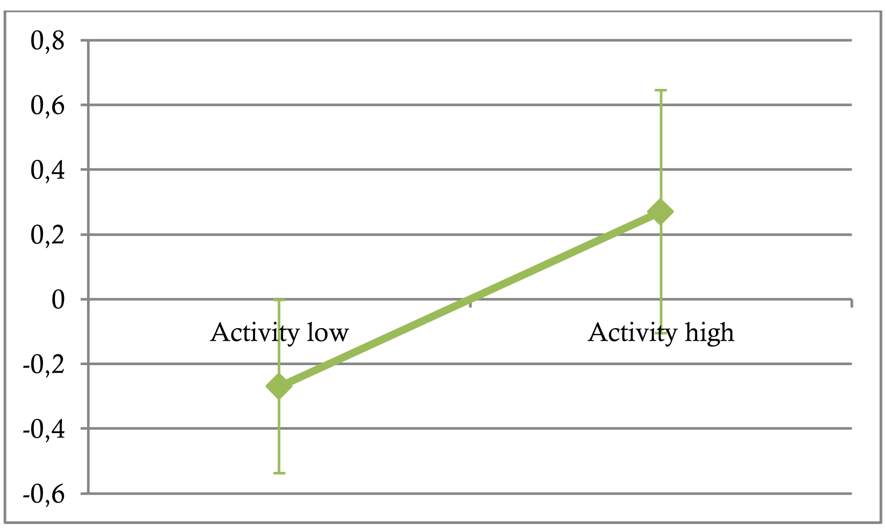

Figure 10. Choice Satisfaction: Activity low vs. Activity high, 1 SE

4.4 Understandability

A significant interaction effect has been found between explanations on/off and the amount of activity on people's understandability of the system, $F(1, 86) = 4.695$, $p = .033$. The mean and standard error of the interaction effect indicate a positive relationship between explanations shown and people's activity. This indicates that people generally understand the system more when they are using the system more frequently and especially when explanations are shown. Explanations seem to contribute to understandability for more active users but not for those who show low activity (figure 11). No significant difference was found between algorithm on or off.

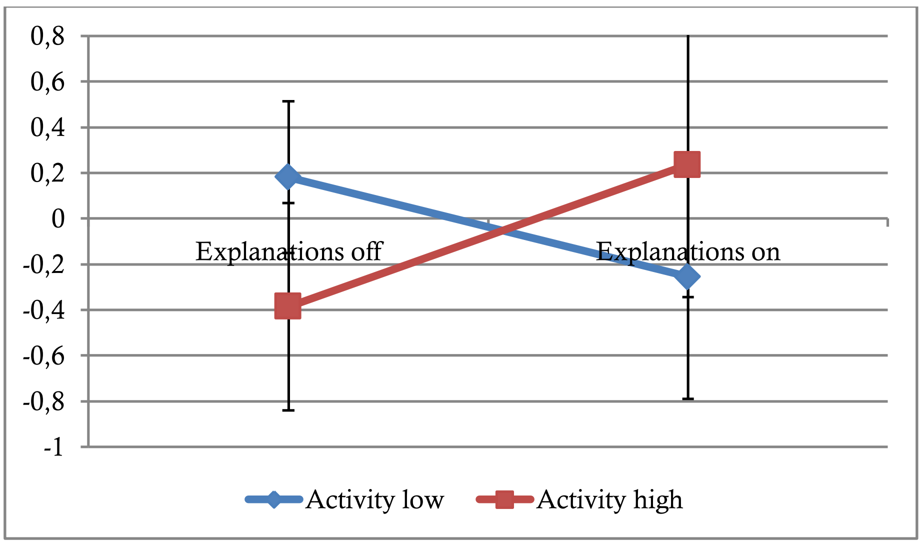

Figure 11. Understandability, 1 SE

4.5 Model

The first attempt to construct a model similar to the one of Knijnenburg et al. (2012) was by structural equation modeling. However, the software (M-plus) that was used to build this kind of model was unsuccessful in actually doing so due to the lack of data. Therefore, another direction was taken. Taking into account all of the significant effects as shown in the previous paragraphs, a model is created that takes into account all of the significant relationships between the different factors. This is done as follows. A t-test is done to show the relationship between the algorithm on and off and the different factors, and the explanations on and off and the different factors. Unfortunately, only a t-test with trust as the testing variable and explanations on and off gave a marginal significant result $t(84) = 1.82$, p

< 0.08. This means that explanations had a positive effect on the trust people had in the system. Using that result, correlations were done between the trust factor and all of the other factors. Trust correlated significantly with four of the factors, being perceived usefulness $r(86) = -.40$, $p < .001$, choice satisfaction $r(86) = .34$, $p < .001$, understandability $r(86) = .31$, $p < .001$ and user interaction satisfaction $r(86) = .64$, $p < .001$. Usefulness correlated negatively and significantly with choice satisfaction $r(86) = -.28$, $p < 0.001$ and user interaction satisfaction $r(86) = -.46$, $p < 0.001$ and also significantly with rating $r(86) = .24$, $p < 0.05$. Choice satisfaction also correlated significantly with understandability $r(86) = .30$, $p < 0.001$, rating $r(86) = .30$, $p < 0.001$ and user interaction satisfaction $r(86) = .38$, $p < 0.001$. Last, but not least, understandability correlated significantly with user interaction satisfaction $r(86) = .28$, $p < 0.001$. Perceived System Quality did not correlate significantly with any of the other factors.

Using these correlations, one could then come up with a model similar to a model that would have been constructed using structural equation modeling. Last, the correlations were checked between all of the factors and the behavior on the actual website. The number of times people viewed the explanations did not correlate significantly with any of the factors. However, the number of times people rated programs did have a significant effect on both usefulness $r(86) = .24$, $p < 0.05$ and choice satisfaction $r(86) = .30$, $p < 0.01$. All of these correlations combined result in the correlation matrix as shown in figure 12. The full figure with all correlations can also be found in Appendix C. The lines between the factors indicate a significant relationship between the two factors.

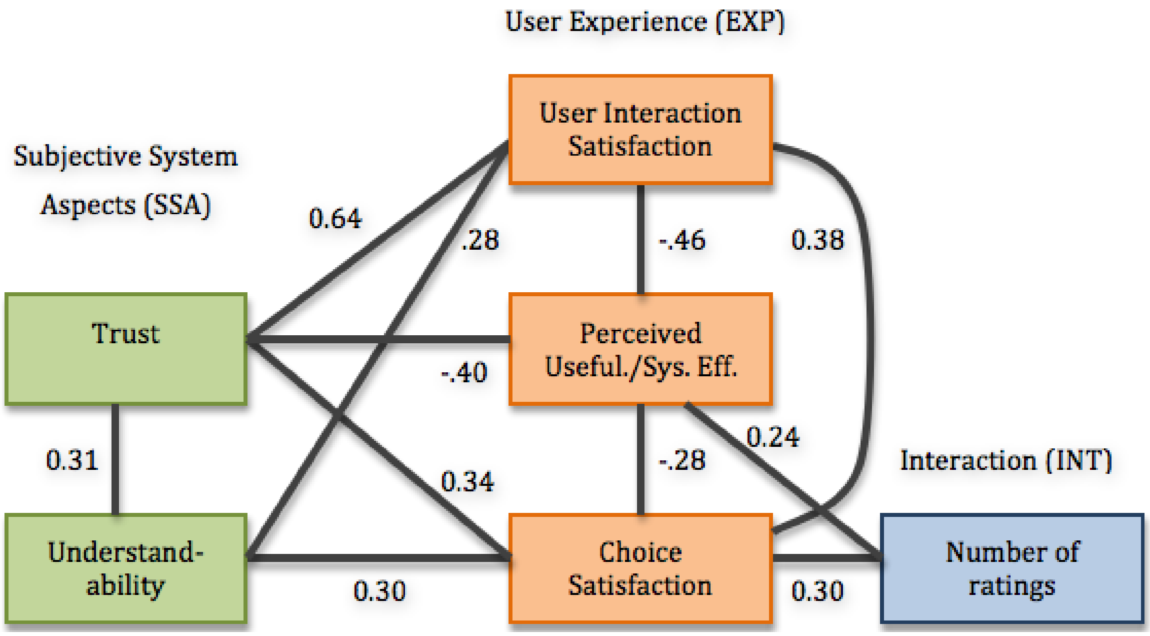

Figure 12. Explanations on/off Factor Correlation Model.

4.6 Summary

After these results, we can revisit the hypotheses as we defined them in chapter 2. Hypothesis 1 stated that content-based recommendations should enhance perceived system quality which increases user interaction satisfaction, choice satisfaction and usefulness. For this study, every ANOVA resulted in a negative effect for content-based recommendations, therefore H1 is simply rejected. User experience and interaction measures do not increase when content-based recommendations are present. Hypothesis 2 stated that explanations should increase trust and understandability, which increases user interaction satisfaction, choice satisfaction and usefulness. Given the results as we presented them in this chapter, this hypothesis can only partially be accepted. Only active users of the system reacted positively on the explanations and the content-based recommendations. Not active users reacted in the exact opposite way. People in the condition with both explanations and content-based recommendations present were generally more negative compared to people in other conditions. In general we found that explanations had a positive effect on the trust people had in the recommendations. This then had a positive effect on understandability, which had a positive effect on the user experience. Hypothesis 3 stated that user experience and interaction measures will be greater for maximizers when explanations are present. H3 was rejected as it was not possible to test. A clear factor could not be found from the factor analysis. Therefore we were not able to distinguish satisficers from maximizers and therefore, there is no statistical proof that user experience and interaction measures are greater for maximizers when explanations are present.

5.0 Discussion

Without taking into account the activity level of participants, the presence of content-based recommendations, explanations, or the combination of the two, has a negative effect on trust, choice satisfaction, and understandability. Results show that people who are in the condition where both explanations and content-based recommendations are present, are generally more negative compared to people in other conditions. Although when the algorithm is turned on, it seem that people are more negative then when explanations are shown. This indicates that participants reacting more negatively to the algorithm rather than about the explanations.

However, when the activity level of participants is taken into account, there is a positive effect on the factors. As participants use the system more frequently, they react more positively to the content-based recommendations and explanations. This suggests a network effect, where the value of a network increases as the number of its nodes increases or in the case of ProgramGenie, the number of user ratings (Hendler & Golbeck, 2008). As participants used the system more, they rated more programs. As they rated more programs, the system had more information to use to fine-tune its recommendations according to the participant's preference. As the participant's preference became more accurate, the explanations can make more sense to the user.

5.1 Limitations

Although sufficient participants were initially recruited, an unforeseen amount of participants dropped out. Of the 78 participants recruited per condition, roughly 20 participants per condition actually participated in the first week and even less participated for both weeks. Also, the uneven amount of participants participating in the first and the second week made it difficult to analyze if there were any differences due to increased familiarity of the system.

The behavioral data obtained from the system also clearly shows that for those participants that did not drop out, only a few participants reached thirty ratings and therefore did not experience the full potential of the system. Almost all participants that were in the condition with explanations viewed the explanations. However, only a few participants saw the explanations more than just a couple of times.

The login rates indicate that people on average logged in two times during the week. Of the 58 people who participated in the first week, only three people logged in more than 6 times, which could indicate that they actually used the system the whole week. It was not possible to check whether participants logged in multiple times on a day or if it is spread over the week. The average login of two times indicates that this is probably too few for the system to get trained and give good recommendations.

Furthermore, participants reported problems with the website. While certain tradeoffs were made to increase system speed, participants still indicated that they had problems reaching the website and that it responded very slowly to their actions. One of the tradeoffs was not showing information of all programs since not every program is offered by every cable company. Therefore, program information for only the most popular TV channels was used since our participants were recruited from all parts of the Netherlands and every region did not receive the same channels. While information of the most important TV channels was provided, some participants reported that the program information as incomplete.

5.2 Conclusion

Previously, Knijnenburg et al. (2012) suggest that SSA in addition to OSA of a recommender system may influence user experience and more specifically, Herlocker and colleagues (2000) as well as Pu and Chen (2007), have found that explanations increase trust and therefore increase user satisfaction on recommender systems. This study was not able to fully replicate the framework by Knijnenburg et al (2012), but did support the idea that trust and choice satisfaction were highly correlated with explanations for recommendations.

5.3 Suggestions for future research

The main limitation in this study is the amount of participants. The number of participants was too low to make an adequate analysis. As for the other limitations in this study, problems with the website could have confounded the results by yielding several non-significant findings. Furthermore, it seems difficult to find committed people for a longer period of time. A better method is needed to get people to be more committed to participate. A future experiment could make use of daily text messages that remind participants to use the system, as well as filling out the questionnaire. For this experiment, we used a raffle among every tenth participant. A more motivational method could be to give every participant a reward or to increase the value of the reward.

References

Herlocker, J.L., Konstan, J.A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. *Proc.CSCW2000*, 241–250.

Iyengar, S. S., & Lepper, M. R. (2000). When Choice is Demotivating: Can One Desire Too Much of a Good Thing? *Journal of Personality and Social Psychology*, 79 995-1006.

Hendler, J. & Golbeck, J.(2008). Metcalfe's law, Web 2.0, and the Semantic Web.*Web Semantics: Science, Services and Agents on the World Wide Web,* 6, (1), 14-20

Knijnenburg, B.P., Reijmer, N.J.M., & Willemsen, M.C. (2011a, in press). Each to His Own: How Different Users Call for Different Interaction Methods in Recommender Systems. *RecSys'11, October 23–27, 2011*, Chicago, Illinois, USA

Knijnenburg, B.P., Willemsen, M.C., & Kobsa, A. (2011b). A Pragmatic Procedure to Support the User-Centric Evaluation of Recommender Systems. *Short paper accepted to the ACM Conference on Recommender Systems (RecSys)*

Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, S., & Newell, C. (2012, in press). Explaining the user experience of recommender systems. *Journal of User Modeling and User-Adapted Interaction*, Vol. 22. Retrieved from http://bit.ly/nuovhh

Ochi, P., Rao, S., Takayama, L., & Nass, C. (2009). Predictors of user perceptions of web recommender systems: How the basis for generating experience and search product recommendations affects user responses. *International Journal Human-Computer Studies,*68, 472–482

Pu, P. & Chen, L. (2006). Trust Building with Explanation Interfaces *IUI'06* 93-100

Pu, P. & Chen, L. (2007). Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems*, 20, 542–556

Sinha, R. & Swearingen, K. (2002). The Role of Transparency in Recommender Systems, *CHI 2002 changing the world, changing ourselves*, 830-831

Appendix A

Questions from Questionnaire

Concept	Question	Question in English
User Persistence	"Ik kijk heel goed naar de eigenschappen van een product om zeker te weten dat het voldoet aan mijn eisen (Oneens - Eens)"	I look at product characteristics carefully to meet my requirements. (Disagree - Agree)
	"Ik ben niet snel tevreden over een product (Oneens - Eens)"	I am not easily satisfied with a product.
Understandability	"Ik begreep goed hoe ik mijn voorkeur kon aangeven en zo de aanbevelingen passender kon maken. (Oneens - Eens)"	I fully understood how I could make my recommendations fit better. (Disagree - Agree)
	"Hoe moeilijk of makkelijk vond je het om je voorkeur aan te geven op de digitale tv gids? (Moeilijk - Makkelijk)"	How easy or hard did you find it to give your preference for the digital TV guide. (Easy - Hard)
Trust	"De digitale tv gids heeft andere belangen. (Oneens - Eens)"	The digital TV guide has other interests. (Disagree - Agree)
	"De digitale tv gids is eerlijk. (Oneens - Eens)"	The digital TV guide is honest. (Disagree - Agree)
	"De digitale tv gids geeft onafhankelijke aanbevelingen (Oneens - Eens)"	The digital TV guide gave independent recommendations. (Disagree - Agree)
	"Ik kan de digitale tv gids volledig vertrouwen. (Oneens - Eens)"	I can fully trust the digital TV guide. (Disagree - Agree)
	"De aanbevelingen van de digitale tv gids zijn betrouwbaar (Oneens - Eens)"	The recommendations of the digital TV guide are trustworthy. (Disagree - Agree)
Perceived System Quality	"Ik vond geen van de aanbevolen items goed (Oneens - Eens)"	I didn't find any good recommendations. (Disagree - Agree)
	"De digitale tv gids voorspelt mijn voorkeur niet accuraat (Oneens - Eens)"	The digital TV guide did not predict my preferences accurately (Disagree - Agree)
	"De aanbevelingen bevatten niet mijn favoriete programma's (Oneens - Eens)"	The recommendations did not consist of my favorite programs. (Disagree - Agree)
Choice Satisfaction	"Ik vond de programma's die ik gekozen heb om te bekijken goed (Oneens - Eens)"	I found the programs that I chose to watch good.(Disagree - Agree)
	"Ik was enthousiast over de items die ik gekozen heb (Oneens - Eens)"	I was enthusiastic about the items I chose.(Disagree - Agree)
	"Ik genoot van het kijken van mijn gekozen items (Oneens - Eens)"	I enjoyed watching my chosen items. (Disagree - Agree)
Perceived usefulness/system effectiveness	"Ik zou de digitale tv gids aanbevelen aan anderen (Oneens - Eens)"	I would recommend the digital TV guide to others.(Disagree - Agree)
	"De digitale tv gids is nuttig (Oneens - Eens)"	The digital TV guide is useful.(Disagree - Agree)

	"Ik kan tijd besparen door gebruik te maken van de digitale tv gids (Oneens - Eens)"	I can save time by making use of the digital TV guide.(Disagree - Agree)
	"De digitale tv gids heeft mij bewuster gemaakt van het aanbod aan programma's. (Oneens - Eens)"	The digital TV guide made me more aware of the programs.(Disagree - Agree)
	"Ik zou de digitale tv gids vaker gebruiken als dat mogelijk was. (Oneens - Eens)"	I would use the digital TV guide more often if it was possible.(Disagree - Agree)
	"Met de digitale tv gids kan ik betere programma keuzes maken. (Oneens - Eens)"	With the digital TV guide, I can make better program choices.(Disagree - Agree)
	"Ik zou de digitale tv gids aan anderen aanraden. (Oneens - Eens)"	I would recommend the digital TV guide to others.(Disagree - Agree)
User Interaction Satisfaction	"Ik vind de digitale tv gids (Vreselijk - Geweldig)	I find digital TV guides (terrible-excellent).
	"Ik vind de digitale tv gids (Moeilijk - Makkelijk)"	I find digital TV guides (easy-hard).
	”Ik vind de digitale tv gids (Frustrerend - Bevredigend)"	I find digital TV guides (frustrating-satisfying).
	Ik vind de organisatie van de digitale tv gids (Verwarrend - Duidelijk)"”	I find the organization of the digital TV guide (clear - unclear)
	"Ik vind de snelheid van de website (Te langzaam - Snel genoeg)"	I find the speed of the website (too slow - fast enough)

Appendix B

Factor analysis of questionnaire results

	1	2	3	4	5
	Trust	Perceived System Quality	Perceived usefulness/system effect	Choice Satisfaction	Understandability
Ik begreep goed hoe ik mijn voorkeur kon aangeven en zo de aanbevelingen passender kon maken.					,942
Hoe moeilijk of makkelijk vond je het om je voorkeur aan te geven op de digitale tv gids?					,867
De digitale tv gids is eerlijk.	,923				
De digitale tv gids geeft onafhankelijke aanbevelingen.	,911				
Ik kan de digitale tv gids volledig vertrouwen.	,804				
De aanbevelingen van de digitale tv gids zijn betrouwbaar.	,921				
De digitale tv gids geeft voldoende informatie die ik nodig heb om een beslissing te nemen	,646				
Ik vond geen van de aanbevolen items goed.		,897			
De digitale tv gids voorspelt mijn voorkeur niet accuraat.		,853			
De aanbevelingen bevatten niet mijn favoriete.		,830			
Ik vond de programma's die ik gekozen heb om te bekijken goed.				,855	
Ik was enthousiast over de items die ik gekozen heb				,909	
Ik genoot van het kijken van mijn gekozen items.				,858	
De items die ik bekeken heb waren een verspilling van mijn tijd		,726			
Ik kan tijd besparen door gebruik te maken van de digitale tv gids			-,701		
De digitale tv gids heeft mij bewuster gemaakt van het aanbod aan programma's.			-,858		
Ik zou de digitale tv gids vaker gebruiken als dat mogelijk was.			-,913		
Met de digitale tv gids kan ik betere programma keuzes maken.			-,922		
Ik zou de digitale tv gids aan anderen aanraden.			-,857		

Appendix C

Correlations between factors

		Trust	Usefulness	Perc. Quality	Choice Sat.	Underst.	Rating	QUIS
Trust	Pearson Corr.	1	-,403**	-,087	,341**	,313**	-,013	,642**
	Sig. (2-tailed)		,000	,424	,001	,003	,903	,000
	N	86	86	86	86	86	86	86
Usefulness	Pearson Corr.	-,403**	1	,055	-,279**	-,128	,240*	-,463**
	Sig. (2-tailed)	,000		,616	,009	,240	,026	,000
	N	86	86	86	86	86	86	86
Perc. Quality	Pearson Corr.	-,087	,055	1	-,004	-,083	,089	-,085
	Sig. (2-tailed)	,424	,616		,969	,447	,413	,434
	N	86	86	86	86	86	86	86
Choice Sat.	Pearson Corr.	,341**	-,279**	-,004	1	,299**	,297**	,375**
	Sig. (2-tailed)	,001	,009	,969		,005	,006	,000
	N	86	86	86	86	86	86	86
Underst.	Pearson Corr.	,313**	-,128	-,083	,299**	1	,135	,282**
	Sig. (2-tailed)	,003	,240	,447	,005		,215	,008
	N	86	86	86	86	86	86	86
Rating	Pearson Corr.	-,013	,240*	,089	,297**	,135	1	-,056
	Sig. (2-tailed)	,903	,026	,413	,006	,215		,593
	N	86	86	86	86	86	94	94
Quis	Pearson Corr.	,642**	-,463**	-,085	,375**	,282**	-,056	1
	Sig. (2-tailed)	,000	,000	,434	,000	,008	,593	
	N	86	86	86	86	86	94	94

**. Correlation is significant at the 0.01 level (2-tailed)   *.Correlation is significant at the 0.05 level (2-tailed)