To Explain or Not to Explain: the Effects of Personal Characteristics When Explaining Feature-based Recommendations in Different Domains

Martijn Millecamp*
Department of Computer Science, KU Leuven
Leuven, Belgium
martijn.millecamp@cs.kuleuven.be

Sidra Naveed*
Department of Computer Science, Duisburg-Essen
University
Germany
sidra.naveed@uni-due.de

Katrien Verbert
Department of Computer Science, KU Leuven
Leuven, Belgium
katrien.verbert@cs.kuleuven.be

Jürgen Ziegler
Department of Computer Science, Duisburg-Essen
University
Germany
juergen.ziegler@uni-due.de

ABSTRACT

In our daily life, we need to sift through various options which often results in choice overload. Recommender systems help to overcome this problem by suggesting potentially relevant items to the users. Explaining the relevancy of these items to users has become an increasingly important goal. In recent years, a large body of research has shown that explanations are an effective means for supporting decision-making processes. However, still little is known on how to best implement these explanations and how these explanations are perceived. In addition, it is unclear how this perception is affected by the product domain or by users' personal characteristics. To fill these research gaps, we conducted an online user study (N=291) with different design mock-ups that represent explanations of feature-based recommendations in various recommendation scenarios in two product domains (music and camera) and using different recommendation techniques (content, collaborative, and hybrid). We conducted in each domain a between-subject study with a baseline without explanations and one of the three designs explaining the feature-based recommendations. The study offers empirical evidence on how the perception of feature-based explanations in various recommendation scenarios are moderated by both the product domains and personal characteristics of the user, in particular need for cognition.

CCS CONCEPTS

• **Human-centered computing → User studies**; *User interface design*; *Visualization design and evaluation methods*; • **Information systems → Personalization**; **Recommender systems**.

KEYWORDS

recommender system; explanations; personal characteristics; evaluation; user characteristics; user modelling; need for cognition

*Both authors contributed equally to this research.

1 INTRODUCTION

The increase of available digital content is improving the chances of finding relevant products which match user preferences and needs, but also introduces a problem of choice overload [27]. To help users overcome this choice overload, recommender systems (RS) such as Netflix, Spotify, Waze, Amazon and Facebook have shown to be helpful and accurate in predicting items that the user might be interested in.

However, the effectiveness of these systems can further be improved from a user perspective, as these systems are sometimes difficult to understand by the user due to their "black box" nature [10, 15]. In this regard, explanations of RS have emerged to be one of the potential solutions to increase the system effectiveness.

More importantly in this context, deciding how to implement these explanations is still difficult [28]. Furthermore, it is still an open question which factors could influence the overall user's perception of the system in the presence of explanations [11, 17].

Although several studies have investigated the effect of explanations, to the best of our knowledge little is known about the (1) effect of explanations on the user's perception for different RS approaches, (2) effect of personal characteristics in terms of decision styles on the perception of different RS approaches, and (3) whether these findings vary in different domains [11, 13, 17].

In this paper, we exploited feature-based data to provide recommendations and explanations for different RS approaches in two different domains and presented these in mock-up designs. We conducted an online user study that investigates the impact of the above mentioned three factors on the assessment of a recommender system. In a within subject design, we compared a baseline without explanations with interfaces explaining three different RS approaches. To investigate the effect of personal characteristics, we analyzed the moderating effect of need for cognition on the perception of the system in both domains. To research whether the results are generalizable across domains, we implemented these interfaces in two domains with a different failure cost and evaluation style: music and digital cameras. The rationale behind these domains is that for a low-risk, experimental domain, such as music, recommender systems have proven to be successful. For high-risk domains with objective features, such as digital cameras, they often have not yet been successful [11, 32].

Based on the open challenges mentioned above, we addressed the following two research questions:

RQ1: Is the overall decision support, choice satisfaction, and perceived recommendation quality dependent on the varying explanations of different RS approaches?

RQ2: Is the overall decision support, choice satisfaction, and perceived recommendation quality with varying explanations of different RS approaches, moderated by the individual's need for cognition? And is this moderating effect dependent on the domain?

The results of our study offer empirical evidence on how decision support (DS), choice satisfaction (CS) and the moderating effect of need for cognition (NFC) are affected by the domains, i.e. music and digital cameras.

2 RELATED WORK

2.1 Explanations in RS

There is an increasing awareness that recommendation effectiveness goes beyond increasing accuracy metrics [10, 30]. Several researchers and practitioners started to use other user-centred metrics such as user satisfaction, diversity and trust to evaluate a RS [30].

In addition, explanations have received more attention, as explaining the rationale behind the recommendation process can increase the acceptance and trust in the system or help to persuade users to follow the recommendation [11].

To explain the rationale to the user, explanations should make clear how the system generated the recommendation. Although there are multiple approaches possible to generate recommendations, we focus on explaining three of the most popular approaches, i.e. content-based, collaborative filtering and the combination of these two (hybrid) [4, 25].

To explain a content-based RS, the system needs to make clear to the user that the recommended items are relevant to the user's feature-based profile. A prominent example of content-based explanations is Tagsplanations, which explains movies based on preferred tags [34]. In our work, the user profile was assumed to be composed of the preferred feature values, with relevance weights for each feature provided by the user [6, 10].

Explaining a RS using collaborative filtering is more challenging as the system needs to justify why users are similar to the active user, as recommendations are generated from similar users' item preferences. An example of such explanations can be found in the PeerChooser system [22]. In this paper, the active user's feature-based profile is used to explain the similarity with other users in terms of item features.

A hybrid system combines multiple approaches, which makes it probably the most complex to explain to the user. A possible solution to explain this approach is to exploit the available information visually, as done in TasteWeights [1]. This interface visualizes why certain music is recommended by a hybrid RS, and also enables users to control the weight of various parameters in the recommendation process. Our hybrid approach combines a collaborative filtering and a content-based approach, by using a feature-based similarity metric which incorporates preferred feature values and the weight of these features [4].

2.2 Effect of personal characteristics on the perception of RS

In recent work, it has been shown that the level and type of information users want to see in a system is generally considered dependent on user's personal characteristics [3]. However, only a limited amount of research investigates, for example, the effect of providing different interfaces for different types of users [9, 12, 13, 18, 20]. In the few studies that did research the effect of personal characteristics on the perception of RS, it was shown that NFC is positively correlated with decision making behavior and that it moderates the user's perception of a RS [18, 33].

In this paper, we investigated the impact of different levels of explanations on the user's perception of the system with people with different NFC, where NFC is a measure of the tendency for an individual to engage in, and enjoy effortful cognitive activities [5].

2.3 Effect of product domains on the perception of RS

Tintarev et al. [32] suggested that the perception of a RS could be influenced by the product domain and they classified domains based on two characteristics: failure cost (cheap vs expensive) and evaluation style (objective vs experimental).

In that regard, they recommend to evaluate a RS in four different domains with the different combinations of characteristics to check whether the results are domain independent or not. In this work, to evaluate our research questions, we considered two domains that differ in regard of two characteristics: the music domain (cheap and experimental) and the camera domain (expensive and objective).

Nevertheless, the effect of different domains on the perception of a RS is only researched to a limited extent.

A work presented by Tintarev et al. [31] evaluated a RS in two domains with a different failure cost, and evaluated the RS in terms of perceived helpfulness. The result of their study showed that users' perception of the system in terms of perceived helpfulness was lower for high failure cost domains compared to low failure domains.

In a more recent work [16], the researchers investigated the differences between domains on the overall assessment of the RS before and after experiencing the recommended item(s). The results indicated that the overall assessment of the RS highly depends on domain as well as type and amount of information provided for the recommendations.

2.4 Effect of non-personalized recommendations on the perception of RS

To evaluate the users' perception of a RS, it is not always needed to give them personalized recommendations. As indicated by [21], if participants believe that the recommendations are coming from a different system, they will react differently. In their study, users were provided with different variants of the system using different recommendation approaches, where in each system the same list of recommendations were provided. The result of the experiment showed that user's perception of the system was different for different approaches, even though in each system users were provided with the same list of items.

Even more, there are some studies showing that some users could benefit from non-personalized recommendations [8, 13, 35]. The study of Knijnenburg et al. [13] researched whether all users are more satisfied with a hybrid RS or whether they preferred a non-personalized recommendations. The result showed that most users liked the hybrid recommendations, but the novice users and maximizers seemed to benefit from the non-personalized recommendations showing only the most popular items. The study presented by Wakeling et al. [35] compared the impact of non-personalized recommendations on the user's performance, search behavior, and system perception in two different application domains. The result of the study revealed that the presence of non-personalized recommendations improved resource discovery, search efficiency and perceived usability. A similar study of [8] compared seven different variants of the system including the non-personalized baseline system, and measured the user' perceived quality of each system, focusing on accuracy, novelty, and overall user's satisfaction. The result of the study showed that simple non-personalized recommendations are well perceived by users in terms of perceived accuracy and overall satisfaction. Benefits in terms of increased sales of non-personalized recommendations have also been shown in the e-commerce application domain [23].

Similar to the above mentioned approaches, we also used a static list of recommendations and present this list in various recommendation scenarios to measure the user's perception of the system.

In contrast to the discussed papers that focus either on different explanations, effects of personal characteristics or effects of different domains, in our study, we combine all of these factors, and measure the impact of different feature-based explanations on the overall user's perception of the system, and investigated whether the results vary in different domains or for users with different need for cognition.

3 VISUAL DESIGN

To investigate our research questions, we followed a similar approach as Ochi et al. [21] and we developed interfaces that were not connected to a real recommender system and provided a static list of items as recommendations. We designed four different interfaces for music as well as the digital cameras domain, each interface representing a different explanation style in terms of recommendation approach. Each recommended item in the interfaces is presented by six different features. In the music domain, similar to the work of [18], these features were selected from the set of audio features of Spotify, based on their popularity (acousticness[1], danceability [2], energy[3], popularity, tempo and valence[4]) [19]. In the camera domain, six common features that are self explanatory, were selected i.e., resolution, continuous shooting, touchscreen, WLAN, camera weight, and battery life.

[1] not having electrical amplification
[2] how suitable a track is for dancing
[3] a perceptual measure of intensity and activity
[4] the musical positiveness conveyed by a track

To enable the comparison between the two domains, the designs were kept uniform between the domains, with only differences in the presented items and their features.

In each domain, a baseline design without explanations was prepared to be compared with three designs that provided explanations for recommendations as they would be generated through content-based filtering (CB), collaborative filtering (CF), and a hybrid technique (HB), by exploiting feature-based data. Each design allowed the user to expand further details of a selected item by clicking on a button. The type and level of information provided in the detailed view depends on the interface and is explained below.

Baseline. In the baseline interface, a detailed view of the selected item is shown (see part a of Figure 1). It provides an image of the item on the left side, a list with the different feature values in the middle and a textual description without any additional explanation on the right.

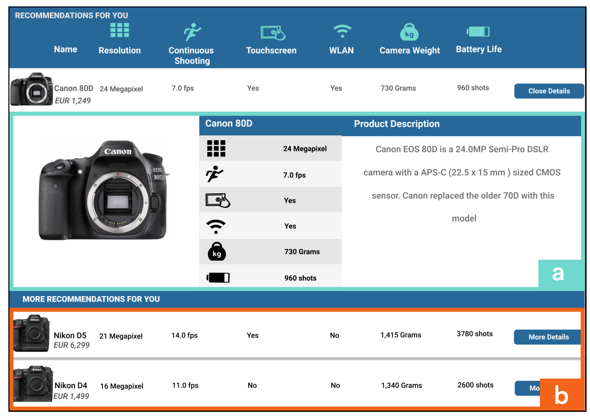


Figure 1: Baseline design: Part a shows the details of the currently selected product. Part b shows other recommendations.

Content-based (CB). To explain a RS that uses a content-based approach, we designed an interface following the guidelines stated by Ribeiro et al. [24]. As mentioned in Section 2.1, to explain this approach, the user needs to be presented with their feature-based profile and the similarity of the recommended items with this profile [6, 25]. In our content-based design, the initial window of the interface is similar to the baseline, but the detailed view of the recommended item as shown in part a of Figure 1 is replaced by Figure 2. The user profile is presented in Figure 2, which shows the user's preferred values of the features with a blue box (Part b) along with the importance of each feature in terms of a 5-point Likert scale (Part a). To show the similarity between the profile of the active user and the recommended item, the features of the current item are indicated with a black line in Part b of Figure 2.

Collaborative filtering (CF). To create a design that explains the recommendations in terms of the collaborative filtering approach, a feature-based similarity metric was exploited to show the similarity of other users with the active user's feature profile. As shown in Part b of Figure 3, for each feature, we show the distribution of preference values of similar users, the preference values of the active user (blue box), and the feature value of the recommended

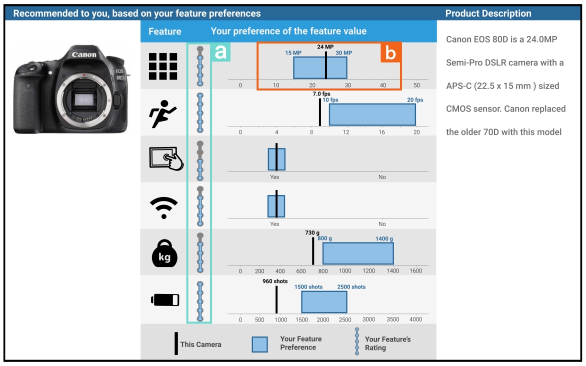


Figure 2: Explanations in the content design: Part a shows the weighting of the features. Part b shows the preferable feature value range of the active user (box) and the value of the current item (line).

item (black line). To explain why the item is recommended, we further show the ratings of this item given by similar users as shown in Part a of Figure 3.

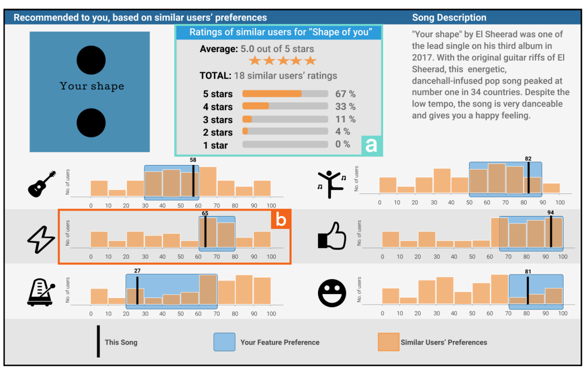


Figure 3: Explanations in the collaborative design of music: Part a shows the ratings of similar users. Part b shows the distribution of preferred values of similar users together with your preferred feature values and the value of the current item.

Hybrid (HB). The third variant of the design explains a RS that combines a content-based approach with collaborative filtering. The hybrid approach is similar to collaborative filtering as it recommends items that similar users liked in the past. Additionally, it also takes into account the importance of each feature for the active user. As such, the visualization of hybrid and collaborative filtering are very similar. The difference between these two interfaces is that for each feature, there is also a 5-point Likert scale to indicate the importance of that feature for the current user as shown in Figure 4 instead of part b in Figure 3.

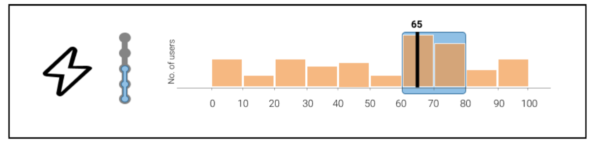


Figure 4: Explanation in the hybrid music design.

4 EXPERIMENTAL DESIGN

4.1 Hypotheses

We conducted a user study to investigate the impact of different types of explanations, product domains, and user's need for cognition, on the overall user's perception of the system. In line with our research questions, we formulated the following hypotheses:

(1) The user's decision support (DS), choice satisfaction (CS), and perceived recommendation quality (PRQ) is dependent on the different types of feature-based explanations.
(2) The user's overall DS, CS, and PRQ with varying levels of feature-based explanations is moderated by the user's need for cognition (NFC), dependent on the product domain.

4.2 Study procedure

To test our hypotheses, we conducted an online user study in which users were randomly assigned to one of the two domains i.e., music or digital cameras. To reduce the effect of individual variance, the users evaluated baseline with either CB, CF or HB design variants in a counter-balanced manner [7]. The three designs i.e., CB, CF or HB were randomly assigned to participants in a between-subject design. In each interface, the user was given the task to pick a relevant song or camera based on the given preferred features. They had as much time as they needed to explore the recommended items and to choose one. Once they decided, they evaluated the interface in terms of DS, CS and PRQ, using a five point likert scale (from disagree to agree), and were presented with another design in which they had to repeat the same task.

4.3 Participants

For the online study, a total of 481 participants were recruited. The majority of the participants were recruited from Amazon Mechanical Turk and were rewarded with $1 as an incentive to complete the study which took 15 to 20 minutes. The other participants were recruited through the university online survey platforms and were rewarded with a half hour credit.

After excluding incomplete responses and outliers, 291 participants (113 females) (Age: M=31.17, SD= 7.50, range 17-62 years) were considered in the analysis. The invalid users were determined based on the time they spent on baseline as well as on the other design variants (CB, CF, HB) and were dependent on the domain. The distribution of the valid participants for both domains is shown in Table 1.

4.4 Measurements

Moderating variables. Before users were presented with the different designs to evaluate, they were asked to provide demographic information and to fill in 18 questions to measure their Need for Cognition (NFC), using the questionnaire by [5]. The NFC was used

Table 1: Distribution of participants between designs for each domain.

	Content	Collaborative	Hybrid
Digital Camera	47	53	44
Music	48	50	49

in our study as a moderating variable (Mean=3.17, Median=3.05, Std.Deviation=0.605) and to test our second hypothesis. A *Kolmogorov-Smirnov test* on both domains indicates that both domains have a similar NFC distribution *D(291) = 0.055, p = 0.977.*

Dependent variables. The impact of different explanation approaches on user's perception of the system is measured, in terms of decision support (DS), choice satisfaction (CS), and perceived recommendation quality (PRQ).

Decision support. One of the major goals of RS is to support users in their decision-making processes [2, 25]. A possible way to support users in their decisions is by providing explanations of recommendations to help users understand the rationale behind the recommendations.

However, the way in which the explanatory information supports users in their decisions is usually not evaluated in RS, which could be measured in terms of several factors i.e, appropriateness and sufficiency of information. For this purpose, we used some questions to measure the impact of explanations on user's decision support, similar to [26] (*Cronbach's alpha=.64*). These questions are as follows[5]:

(1) The information provided for the recommended items was easy to understand.
(2) The information was helpful for deciding which item(s) to select.
(3) The amount of information provided for each recommended item was just right.
(4) I would like to see more information for the recommended items.*
(5) I would like to see less information for the recommended items.*
(6) The type of information helped me to make my decision.
(7) The information provided helped me to decide quickly.
(8) Overall, I found it difficult to decide which items(s) to select.*

Choice satisfaction. To measure if the domain, the type and amount of information influences how satisfied the users are with their choice, we used three questions from the framework of Knijnenburg et al. to measure the choice satisfaction (CS) [14]. These questions are as follows:

(1) I think I chose the best item from the options.
(2) I would recommend the chosen item to others.
(3) I am satisfied with the item I have chosen.

Perceived recommendation quality. Even if the recommendation quality was the same for all designs, we wanted to measure if the users also perceived it in that way as they see different types and amount of information. We used three questions from Knijnenburg et al. [14] to measure the perceived recommendation quality (PRQ) and are as follows:

(1) The shown items fitted my preferences.
(2) Overall, the shown items were well-chosen.
(3) Overall, the system showed me too many bad items*.

[5]Questions with a * are reverse scored

Table 2: Significant differences with the baseline

Domain	Design	DV	Diff of Means	SD	p	t
Digital Cameras	Content	DS	-0.16	0.46	.021	2.38
	Content	CS	-0.23	0.63	.015	2.52
	Collaborative	PRQ	-0.2	0.71	.045	2.05
Music	Collaborative	DS	0.17	0.48	.013	-2.58

5 RESULTS

5.1 Comparison with the baseline

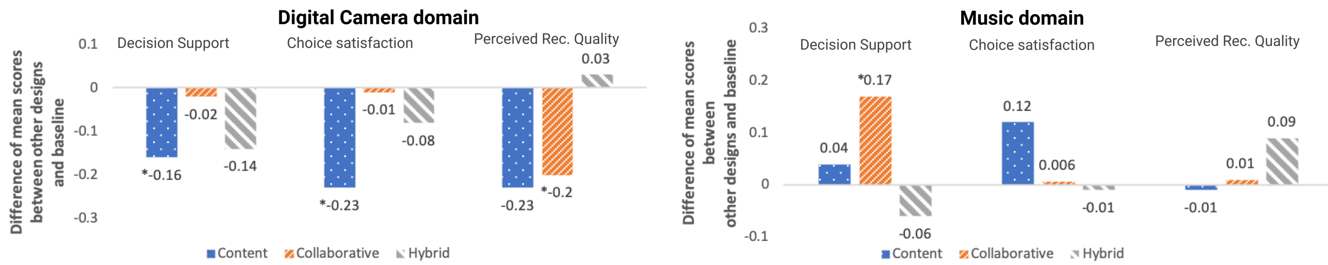
General. Users were presented with the baseline design as well as one of the three designs for content-based (CB), collaborative filtering (CF), and hybrid (HB) recommendations in a within-subject study using a randomized order. Before we investigated the perception of each design in comparison with the baseline design, we tested for the order effect of the baseline with the CB, CF, and HB, with MANOVA ($\alpha = 0.05$) in both domains, but there was no significant order effect. This shows that the order of the baseline and the other design does not impact the user's perception of the system, and therefore, we did not consider the order of designs in our further analysis.

Comparing the domains. We further investigated the users' perception of the designs in comparison with the baseline, in both the digital cameras and music domain. To reduce the personal bias effect, we took a difference of the mean scores of CB, CF, and HB with the baseline, and compared the designs in terms of DS, CS, and PRQ. The result can be seen in Figure 5a and 5b, where the positive value indicates that the CB,CF or HB design scored better than the baseline for the respective dependent variable. We ran multiple paired sample t-tests for each domain separately to see for which dependent variables the differences are significant. The significant results are listed in Table 2 and show that in the digital cameras domain, the baseline design significantly performed better than content design in terms of DS and CS. For PRQ, the performance of baseline design was significantly better than the collaborative design. In the music domain, the collaborative design was significantly better than the baseline in terms of DS.

5.2 Testing Hypothesis 1

In our first hypothesis, we wanted to verify if the perception of a RS is dependent on different types of explanations. To test this, we used MANOVA on the difference in the mean scores of CB, CF, HB, with the baseline [6]. The results revealed statistically insignificant difference for the digital cameras $F(6, 278) = 1.53, p = 0.16; Wilk's\ \lambda = 0.93, \eta^2 = 0.03.$ and music domain $F(6, 284) =$

[6]It has been shown in [29] that parametric test like MANOVA can be used for the ordinal data types. In this and all other analysis, an alpha value of 0.05 is used.



(a) Comparison of the performance for digital camera domain

(b) Comparison of the performance for music domain

Figure 5: Comparison of the performance of the baseline design with content, collaborative, and hybrid designs for Digital Camera and Music domain (* indicating the significant difference, where the positive value indicates that the CB,CF, or HB designs scored better than the baseline design)

$1.18, p = 0.31; Wilk's\ \lambda = 0.95, \eta^2 = 0.02$, for aggregated dependent variables. This shows that user's perception of DS, CS, and PRQ is not dependent on varying types of feature-based explanations in terms of CB, CF, and HB styles in both digital cameras as well as in the music domain.

Comparing the domains. To further compare the means of DV between domains, we applied an independent t-test and we found 3 significant results, shown in Table 3. For DS, we see that the content and the collaborative design are performing significantly better in the music domain than in the camera domain. For CS, we see that the content design is performing significantly better in the music domain than in the camera domain.

Table 3: Significant differences between product domains in terms of DS, CS, and PRQ

	Condition	Domain	Mean	SD	t	p
Decision Support	Content	Digital Camera	-.16	.46	-1.98	.050
		Music	.04	.54		
	Collaborative	Digital Camera	-.02	.55	-1.99	.049
		Music	.17	.55		
Choice Satisfaction	Content	Digital Camera	-.23	.63	-3.1	.002
		Music	.12	.45		

5.3 Testing Hypothesis 2

In our second hypothesis, we wanted to verify if the perception of a RS is moderated by the user's NFC. In order to test the second hypothesis, we used MANOVA to seek the moderating effect of NFC. We hypothesized that the moderating effect of NFC could be dependent on product domains. Therefore, we conducted the analysis on both domains separately.

Digital Cameras domain. The result of multivariate test for the digital cameras domain revealed no significant interaction effect between the different feature-based explanations (CB, CF, HB) and NFC, $F(84, 189) = 0.92, p = 0.65; Wilk's\ \lambda = 0.35, \eta^2 = 0.29$. This indicates that the NFC does not moderate the relationship between the designs (CB,CF, HB) and the user's perception of the system in terms of DS, CS, and PRQ.

Music domain. In contrast to the digital cameras domain, there was a significant interaction effect between different feature-based explanations (CB, CF, HB) and NFC in the music domain $F(99, 198) = 1.49, p = 0.009; Wilk's\ \lambda = 0.18, \eta^2 = 0.42$.

This interaction effect indicates that the NFC moderates the relationship between CB, CF, HB and dependent variables in the music domain as can be seen in Figure 6.

To investigate which relationship between design and DV was significantly influenced by the NFC, we looked at the between-subjects effects and found that DS (p=0.002) and CS (p=0.022) were significant. This shows that DS and CS were responsible for the overall significant interaction effect of NFC.

The insignificant interaction effect in the digital cameras domain and the significant interaction effect in the music domain indicate that NFC moderates the relationship between our different feature-based explanations (CB, CF, HB) and dependent variables, but the moderation is dependent on the product domain, thus validating Hypothesis 2.

5.4 Analyzing the moderating effect in music domain

Nature of the moderating effect in music domain. The results of the MANOVA analysis in section 5.4, shows that NFC moderates the relationship between feature-based explanations in terms of CB, CF, HB styles and DV for the music domain. However, it does not explain the trend of the moderating effect, i.e., how low and high NFC values moderate the relationship between feature-based explanations (CB, CF, HB) and dependent variables in the music domain. To investigate this pattern of results, we did a median split on the NFC to determine the low and high NFC behavior, and plotted line graphs for each dependent variable. For the music domain, we can see the trend of moderating effect in terms of low and high NFC values in figure 7. In general, the performance of the content design for high NFC is always higher than the low NFC, across all three dependent variables. However, different behavior can be observed for low and high NFC in collaborative and hybrid designs, across the three dependent variables.

Decision support. In terms of DS, the collaborative design performed better than the other two designs for both high and low NFC individuals. For collaborative and hybrid designs, the DS is better for low NFC as compared to the high NFC individuals. However, an opposite trend can be seen for the content design. Interestingly in this context, an interaction effect can be observed between the content and hybrid designs.

Choice satisfaction. In terms of CS, the content design performed better than the other two designs. However, high NFC individuals seem to have better CS for the content design than low NFC individuals. The same trend can be observed for the hybrid design too, but for the collaborative design, no visual moderating effect can be observed.

Perceived recommendation quality. For PRQ, we can observe some moderating effects for all three designs, but this effect is relatively higher for the content design than for the other two designs, where PRQ is significantly better for individuals with high NFC than individuals with low NFC.

6 DISCUSSION

The results of the user study showed three main interesting findings: (1) The insignificant difference in the user's perception of varying feature-based explanations in terms of DS, CS, PRQ; (2) The effect of different domains on user's perception of varying feature-based explanations in terms of DS, CS, and PRQ; (3) The user's perception of the system in the presence of varying feature-based explanations is moderated by the need for cognition (NFC), which is dependent on the product domains.

6.1 Difference in varying feature-based explanations

As discussed in Section 5.2, we analyzed the effect of varying feature-based explanations on DS, CS and PRQ.

The results of a MANOVA test revealed that the difference between the CB, CF, and HB were insignificant for the aggregated dependent variables (DS, CS, PRQ), which lead to rejecting our first hypothesis. This is a contradictory result to the findings of Ochi et al. [21] in which they showed a difference in perception between two designs, even when the systems shows the static non-personalized recommendations. Even though we can see some differences between the user's perception of the baseline and the other three interfaces (CB, CF, HB) in Figure 5 and Table 2. A possible explanation for the rejection of our hypothesis 1 could be that in our case the interfaces CB, CF and HB are too closely related to each other to be perceived significantly different by the users.

6.2 Difference in the users' perception of the system between domains

As discussed before, we noticed some differences between the user's perception of the baseline and CB, CF and HB. From Figures 5a and 5b and in Table 2, it is clear that these differences are different in the music and the camera domain. In the camera domain, the baseline is perceived slightly better in almost all cases, while in the music domain no clear trend can be observed in terms of users' perception of designs. We assume that the reason for this difference can be twofold.

The first reason for this result could be that there are more novice users in the cameras domain than in the music domain. Previous work has already shown that novice users seem to benefit from a non-personalized recommender with only the most popular items [13].

The second reason for the better performance of the baseline in the cameras domain could be the high failure-cost associated with digital cameras, as it has already been shown that failure-cost could impact the way users make their decisions [26]. The high failure-cost might have forced users to think rationally and rely more on their judgments rather than on systems with more complex explanations which is also visible in the moderating effect of NFC, discussed in the next section. This reason could be justified with the study presented in [20], where the result of the study showed that in the digital cameras domain, the rational thinkers perceive the system independent of the level of explanations provided to them. This results in users not relying on recommendations and explanations for making their decisions.

6.3 Difference in the effect of need for cognition on the domain perception

Our results further shows that the user's perception of the system in the presence of feature-based explanations is moderated by an individual's need for cognition (NFC), dependent on the product domain. The result shows that NFC is not moderating the user's perception in the digital cameras domain, but is moderating in the music domain, as shown in Figure 6. The result might again be explained in terms of difference in evaluation style, failure-cost and experience. The more objective nature of the features, the higher failure-cost and the lack of experience in the digital camera domain might have forced all users to invest more cognitive effort to make the decision, driving them to behave like persons with a high NFC. This assumption is supported by the moderating effect of NFC that is present in the music domain in which users are possibly more free to spend the amount of cognitive effort they want. This could implicate that for novice users, making a decision with a high failure-cost based on objective features overwrites their personal characteristics to involve oneself more or less cognitively in a task.

The nature of the moderating effect in music domain. As shown in Figure 7, for the content design, a similar moderating trend of NFC can be observed across DS, CS, and PRQ. The trend shows that the content design performed better for high NFC than for low NFC. A possible reason could be that the content design is only providing information about features of an item, which may be sufficient for people with high NFC to make their decisions. This might result in better perception of the system in terms of DS, CS, and PRQ. Low NFC individuals are not willing to put a cognitive effort in understanding the information, thus resulting in a lower perception of the system.

In contrast to the content design, the moderating effect of NFC on the collaborative and hybrid is not consistent. For DS, both designs scored better for low NFC than for high NFC, which is the opposite in the case of PRQ. For CS, both designs showed the opposite trend for low and high NFC. This inconsistency in the moderating effect of NFC for collaborative and hybrid designs could be explained by

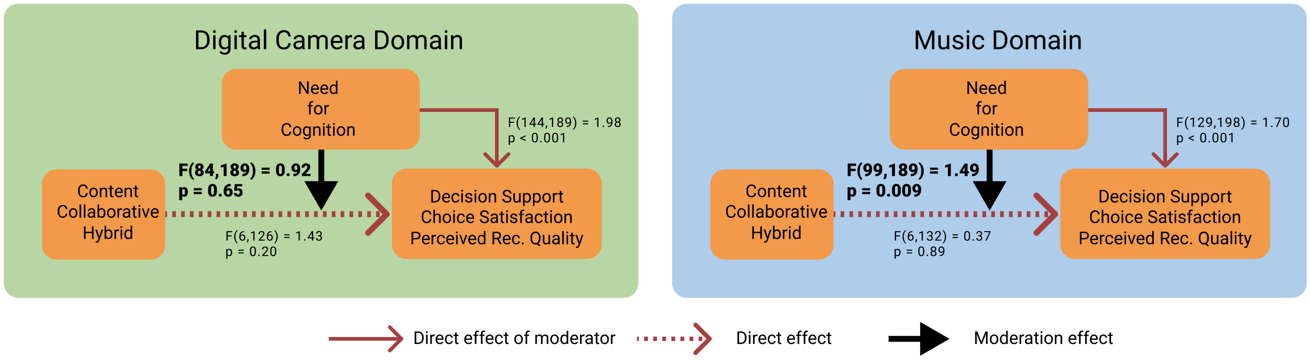


Figure 6: Moderating effect in digital cameras and music domain

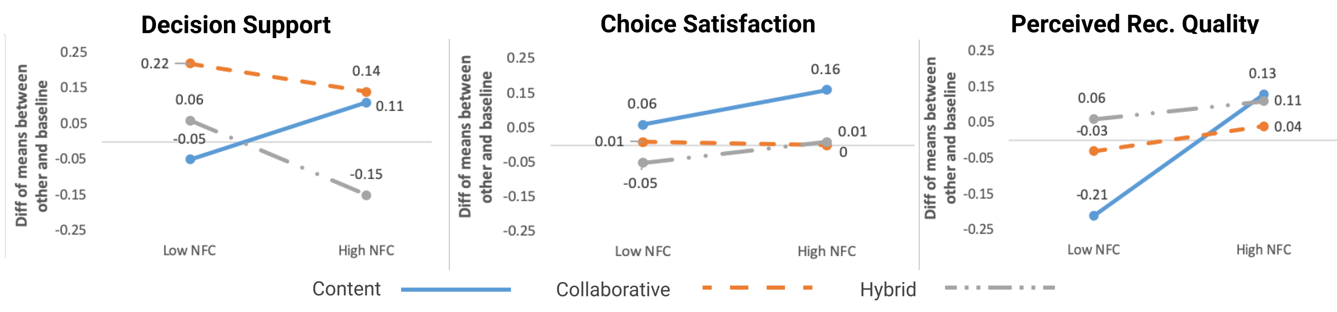


Figure 7: Moderating effect in music domain

several potential factors, such as complexity of designs, similarity of designs in terms of presented feature-based information, etc.

7 CONCLUSION AND FUTURE WORK

In this paper, we have investigated the impact of different feature-based explanations on user's perception of a RS through an extensive empirical study. We found that the overall user's perception of the system does not depend on different feature-based explanations, however, there are other factors that influence the users' perception of the system: difference in domains and difference in Need for Cognition. We investigated these factors to some extent, and the result of our study showed some interesting findings.

Overall, in the digital cameras domain, users preferred a baseline design of the system without explanations. We found an opposite trend in the music domain, where users seem to be preferring versions of the system with explanations.

Additionally, we found that the user's perception of the system is moderated by the user's Need for Cognition in the music domain only. A potential candidate for explaining this difference in moderating effect between domains could be that the more objective nature of the features, the high failure-cost and the lack of experience in the digital cameras domain might have forced all users to invest more cognitive effort to make the decision, driving them to behave like persons with a high NFC. In the music domain there are less novice users, the features are evaluated more experimentally and decisions have a lower failure-cost. Thus the user's personal characteristics are more prominent, and still maintaining the user's personal impact on the overall system perception. This could implicate that for novice users, making a decision with a high failure-cost based on objective features overwrites their personal characteristics to involve oneself more or less cognitively in a task.

Even though we showed the influence of product domains on user's overall perception of the system, we cannot verify which domain factors have caused these differences. This might be due to the difference in the evaluation style and failure-cost, which characterizes the domain, but also due to several other potential factors e.g. experience with the domain, complexity and physicality of the product, etc. These factors needs to be explored and addressed further in the future work. In future work, we plan to focus on methods to implement various RS approaches, in multiple domains to verify the results of this paper. This would also enable us to research in more detail the way explanations should be presented in different domains and the effect of personal characteristics on this.

REFERENCES

[1] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. TasteWeights: a visual interactive hybrid recommender system. In *Proc. of RecSys'12*. ACM, 35–42.

[2] Brand, Christian Laier, Mirko Pawlikowski, and Hans J. Markowitsch. 2009. Decision making with and without feedback: The role of intelligence, strategies, executive functions, and cognitive styles. *Journal of Clinical and Experimental Neuropsychology* 31, 8 (2009), 984–998.

[3] Matthias Brand, Christian Laier, Mirko Pawlikowski, and Hans J Markowitsch. 2009. Decision making with and without feedback: The role of intelligence, strategies, executive functions, and cognitive styles. *Journal of Clinical and Experimental Neuropsychology* 31, 8 (2009), 984–998.

[4] Robin Burke. 2007. Hybrid web recommender systems. In *The adaptive web*. Springer, 377–408.

[5] John T Cacioppo and Richard E Petty. 1982. The need for cognition. *Journal of personality and social psychology* 42, 1 (1982), 116.

[6] Jorge Castro, Rosa M Rodriguez, and Manuel J Barranco. 2014. Weighting of features in content-based filtering with entropy and dependence measures. *International journal of computational intelligence systems* 7, 1 (2014), 80–89.

[7] Gary Charness, Uri Gneezy, and Michael A Kuhn. 2012. Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization* 81, 1 (2012), 1–8.

[8] Paolo Cremonesi, Franca Garzotto, Sara Negro, Alessandro Vittorio Papadopoulos, and Roberto Turrin. 2011. Looking for âĂIJgoodâĂİ recommendations: A comparative evaluation of recommender systems. In *IFIP Conference on Human-Computer Interaction*. Springer, 152–168.

[9] Alexander Felfernig, Gerhard Friedrich, Dietmar Jannach, and Markus Zanker. 2006. An integrated environment for the development of knowledge-based recommender applications. *International Journal of Electronic Commerce* 11, 2 (2006), 11–34.

[10] Chen He, Denis Parra, and Katrien Verbert. 2016. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications* 56 (2016), 9–27.

[11] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 241–250.

[12] Michael Jugovac, Ingrid Nunes, and Dietmar Jannach. 2018. Investigating the decision-making behavior of maximizers and satisficers in the presence of recommendations. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*. ACM, 279–283.

[13] Bart P Knijnenburg, Niels JM Reijmer, and Martijn C Willemsen. 2011. Each to his own: how different users call for different interaction methods in recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. ACM, 141–148.

[14] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 441–504.

[15] Benedikt Loepp, Catalin-Mihai Barbu, and Jürgen Ziegler. 2016. Interactive Recommending: Framework, State of Research and Future Challenges.. In *EnCHIReS@ EICS*. 3–13.

[16] Benedikt Loepp, Tim Donkers, Timm Kleemann, and Jürgen Ziegler. 2018. Impact of item consumption on assessment of recommendations in user studies. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 49–53.

[17] Alex C Michalos. 1970. The costs of decision-making. *Public Choice* 9, 1 (1970), 39–51.

[18] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2019. To Explain or not to Explain: the Effects of Personal Characteristics when Explaining Music Recommendations. In *Proceedings of the 2019 Conference on Intelligent User Interface. ACM*. 1–12.

[19] Martijn Millecamp, Nyi Nyi Htun, Yucheng Jin, and Katrien Verbert. 2018. Controlling Spotify recommendations: effects of personal characteristics on music recommender user Interfaces. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*. ACM, 101–109.

[20] Sidra Naveed, Tim Donkers, and Jürgen Ziegler. 2018. Argumentation-Based Explanations in Recommender Systems: Conceptual Framework and Empirical Results. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. ACM, 293–298.

[21] Paloma Ochi, Shailendra Rao, Leila Takayama, and Clifford Nass. 2010. Predictors of user perceptions of web recommender systems: How the basis for generating experience and search product recommendations affects user responses. *International Journal of Human-Computer Studies* 68, 8 (2010), 472–482.

[22] John O'Donovan, Barry Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. 2008. PeerChooser: visual interactive recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1085–1088.

[23] Bhavik Pathak, Robert Garfinkel, Ram D Gopal, Rajkumar Venkatesan, and Fang Yin. 2010. Empirical analysis of the impact of recommender systems on sales. *Journal of Management Information Systems* 27, 2 (2010), 159–188.

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1135–1144.

[25] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. Introduction to recommender systems handbook. In *Recommender systems handbook*. Springer, 1–35.

[26] María Luisa Sanz de Acedo Lizarraga, María Teresa Sanz de Acedo Baquedano, Maria Soria Oliver, and Antonio Closas. 2009. Development and validation of a decision-making questionnaire. *British Journal of Guidance & Counselling* 37, 3 (2009), 357–373.

[27] Barry Schwartz. 2004. *The paradox of choice: Why more is less*. Vol. 6. HarperCollins New York.

[28] Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 107–120.

[29] Gail M Sullivan and Anthony R Artino Jr. 2013. Analyzing and interpreting data from Likert-type scales. *Journal of graduate medical education* 5, 4 (2013), 541–542.

[30] Nava Tintarev and Judith Masthoff. 2007. A survey of explanations in recommender systems. In *2007 IEEE 23rd international conference on data engineering workshop*. IEEE, 801–810.

[31] Nava Tintarev and Judith Masthoff. 2008. Over-and underestimation in different product domains. In *Workshop on Recommender Systems associated with ECAI*. 14–19.

[32] Nava Tintarev and Judith Masthoff. 2009. Evaluating recommender explanations: Problems experienced and lessons learned for evaluation of adaptive systems. In *In the workshop on User-Centred Design and Evaluation of Adaptive Systems in association with UMAP'09*. Citeseer.

[33] Stephanie Tom Tong, Elena F Corriero, Robert G Matheny, and Jeffrey T Hancock. 2018. Online Daters' Willingness to Use Recommender Technology for Mate Selection Decisions. In *Proceedings of the 5th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with ACM Conference on Recommender Systems (RecSys 2018)*. ACM, 45–52.

[34] Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on Intelligent user interfaces*. ACM, 47–56.

[35] Simon Wakeling, Paul Clough, and Barbara Sen. 2014. Investigating the potential impact of non-personalized recommendations in the OPAC: Amazon vs. WorldCat.org. In *Proceedings of the 5th Information Interaction in Context Symposium*. ACM, 96–105.