Do you come here often?

investigating the effects of domain knowledge and controllability on a theme park recommender system

Smits, T.H.J.

Award date:
2021

Eindhoven, 22-1-2021

Do you come here often?

Investigating the effects of domain knowledge and controllability on a theme park recommender system.

By T.H.J. Smits

Identity number: 0889679

In partial fulfilment of the requirements for the degree of

Master of Science

in Human-Technology Interaction

Supervisors:

Dr. ir. M.C. Willemsen

Dr. W. Barendregt

Company Supervisor:

J. Rietbergen

ABSTRACT

To assist visitors in getting even more out of their visit, Efteling has launched a conversational recommender system to provide suggestions for attractions to visit in their theme park. Literature suggests that the optimal level of control over indicating user preferences may be dependent on the user's knowledge about the decision domain. In order to investigate this for visitors of a theme park, the controllability of the recommender system was manipulated in a series of A/B tests. By implementing different preference elicitation methods and varying the number of items in the initial questionnaire, the effects of controllability on recommendation satisfaction and persuasiveness are researched. The findings seem to suggest that when users perceived a low level of control due to lacking the domain-specific knowledge to answer a question, this negatively affected their recommendation satisfaction. The level of domain-specific knowledge also affected the preferred level of domain-specific control that the user was offered. However, further research is highly encouraged.

Acknowledgements

First and foremost, I would like to express my gratitude to Martijn Willemsen for his invaluable involvement in this thesis. Even though, for the complete duration of the project, it was not possible to have a meeting in the same room, he was able to share helpful insights, motivating words, and an inspiring passion for the topic. Additionally, I would like to thank my second supervisor, Wolmet Barendregt, for her useful perceptions that have brought this thesis to a higher level.

Secondly, I thank the app and guest data team at Efteling, and Jonas Rietbergen in particular, for granting me the opportunity to contribute to the recommender system and for making me feel welcome in the team, both at the office and in digital meetings. It was a true pleasure working at Efteling. The freedom they provided, as well as the lively discussions and critical questions, were truly helpful in shaping the project and the study.

I also want to thank my friends and my family for their support throughout this project in particular, and my educational career in general. I feel fortunate to have them around me. Last, but most certainly not least, I would like to thank my girlfriend, Simone. Thank you for your endless support and motivation throughout this project and in life, and most importantly, your willingness to spend a lockdown with me.

Table of Contents

Introduction

Efteling

Having welcomed 5.4 million visitors in 2019, Efteling, situated in the south of the Netherlands, is the third largest theme park in Europe (Themed Entertainment Association & AECOM, 2019). The theme park features a number of attractions for all ages and its fairy tale theme makes it a popular destination for family visits. In their vision for 2030, Efteling has expressed their ambition to be the first theme park resort in Europe to receive a rating of 9+ for every visiting group (Efteling, 2019). To achieve this goal, five domains have been highlighted in which clear potential for improvement has been described. One of these domains is the use of technology. Efteling is aware of the role that innovative use of technologies can play in improving the overall experience of their visitors.

One of the technological domains that Efteling will explore in pursuit of a 9+ rating is that of recommender systems. By providing guests with tailor-made suggestions of what attractions to visit next, based on, for example, user preferences and location data, a recommender system can help guests of the park to get more out of their day. The study presented in this thesis investigates aspects of human-recommender interaction in this unique domain.

Recommender systems

Recommender systems are decision-making tools that are designed to assist users in domains where they lack a complete understanding of all available alternatives (Resnick & Varian, 1997). In a similar way to how people would ask other (potentially more knowledgeable) people for recommendations when making certain decisions, a recommender system can also provide these suggestions. Recommender systems may be useful for people who are generally unfamiliar with the decision items, but they can also be used to narrow down large item sets or consider aspects of the decision that may not be readily available to the user.

Fundamentally, the main task of a recommender system is predicting which items in a set are most suited to be recommended to a specific user. These predictions can be made by calculating a utility score for each item, based on an algorithm. Multiple approaches exist that can be implemented and combined to optimize this algorithm. The two most prominent techniques that are often used are collaborative filtering and a content-based approach (Burke, 2007; Adomavicius & Tuzhilin, 2005). Collaborative filtering analyzes ratings that the user has given in the past and identifies users with similar rating histories. Based on these peers, it makes predictions about the suitability of items that the user

has not yet rated (Burke, 2007). Recommender systems that use a content-based approach will recommend items that are similar to items that the user has rated positively in the past, based on a number of specific item features (Adomavicius & Tuzhilin, 2005). Often, multiple approaches are combined in a so-called hybrid approach, in which two or more approaches complement each other to achieve higher prediction accuracy (e.g. Balabanović & Shoham, 1997). Examples of other recommendation approaches are demographic (i.e. comparing items based on ratings of users with similar demographic features) and knowledge-based (i.e. using specific knowledge about the decision domain to find items that might fulfill the needs of the user) (Burke, 2007). The recommender system developed by Efteling features a hybrid approach, recommending attractions based on the experiences of similar users, combined with item features of the attractions themselves, such as location and waiting times.

Traditionally, the development of recommender systems was mainly concerned with developing algorithms that achieve a prediction accuracy that is as high as possible and recommender systems were also evaluated by these metrics (Adomavicius & Tuzhilin, 2005). Recent years, however, mark a shift in approach, as increasingly more emphasis is put on user-centric factors, such as visualizing the recommendations attractively and ensuring that the system is satisfactory for the user (He, Parra, & Verbert, 2016). This shift has gone hand in hand with a similar shift in the evaluation of recommender systems and research in the domain, which also started to embrace the notion that a recommender system is only truly useful if the target group actually wants to use it (Pu, Chen & Hu, 2012; Knijnenburg, Willemsen, Gantner, Soncu, & Newell, 2012).

An important facet of designing a recommender system that people actually want to use is the way users interact with it. It is important that users receive satisfactory recommendations that are also relevant to the context of the request. In addition to this, a recommender system needs to learn the preferences of the user to predict the most suitable recommendations. The collecting of this information also needs to be designed in a non-obtrusive way. A multitude of different platforms and user interfaces has been implemented in the past and designing a method of interacting that is contextually relevant depends heavily on the decision domain (Lu, Wu, Mao, Wang, & Zhang, 2015). In this thesis, a recommender system is discussed that is deployed in a theme park to assist tourists visiting Efteling. This domain contains a relatively small item space, with several contextual aspects influencing the suitability of the items for a certain user, such as waiting times and the location of the user in the park. Efteling has developed a mobile recommender system in the form of a smartphone application. A mobile design was chosen because of the dependence on location, the relevance of an item fluctuating over time, and the fact that a user may request recommendations multiple times during their visit. Efteling aims to provide a fun and intuitive user experience, in which the whimsical nature of the theme

park shines through. Therefore, they required a design that allows for straightforward interaction in a manner that is similar to how a visitor might receive recommendations from an employee in the park. Conversational-style recommender systems are suitable for this purpose, as they are designed to resemble natural interactions between people (Christakopoulou, Radlinski, & Hofmann, 2016). For this reason, Efteling has chosen this style of recommender systems, which has been associated with the domain of tourism before (Gretzel, 2011). As such, the current thesis is concerned with conversational-style recommender systems.

Conversational-style recommender systems

Users of a conversational-style recommender system engage in a chat-based conversation with a virtual agent. This method of recommending is argued to relatively closely resemble the way humans provide recommendations to each other, as it simulates the social act of having a conversation (Christakopoulou et al., 2016). During this conversation, by asking questions, the system gathers information about relevant user-specific aspects, and it proposes items that match the user's profile. To establish this profile and to provide personalized recommendations, information needs to be collected to distinguish the user from others. In this section, two important ways of collecting information about the user are discussed: the initial set of questions before providing the first recommendation, and critiquing.

The initial set of questions

When a new user starts using a recommender system, this can lead to challenges that are commonly referred to as cold start problems, as was highlighted by Bobadilla, Ortega, Hernando, and Bernal (2012). They stated that collecting additional information about the user is an effective way of providing personalized recommendations to new users. While initializing a user profile, a recommender system may ask a number of questions, inquiring about the preferences of the user. Using this information, a profile of the user is built, based on which the system can start providing personalized recommendations. Several kinds of Preference Elicitation (PE) methods exist, a few highlighted below, which can be used to extract these user aspects.

Attribute-based PE is a popular PE method that allows users to indicate their preferences and their priorities about specific attributes of the items in the decision domain (e.g. Häubl & Trifts, 2000). This method is particularly useful for users who have concrete demands for the product of choice, and want to filter a large item set, based on certain criteria. For example, when looking for a new television, a user might want to indicate preferences regarding price, resolution, dimensions, and other specific aspects of the large item space, to reduce the number of options to manually compare. This does, however, require a certain level of knowledge about the various items to be chosen from. According to

Hutton and Klein (1999), people with relatively low domain knowledge have more difficulty breaking down a concept into specific attributes and making a comparison between them.

Needs-based PE is a different form of preference elicitation that focuses on the desired outcome of choosing a certain item, rather than on the items themselves. Users who are less experienced with the domain in which they are choosing, might not be knowledgeable enough to indicate explicit attribute preferences, as argued by Randall, Terwiesch, and Ulrich (2007). For this reason, they developed a recommender system in which users indicated their personal needs regarding the decision domain. The system associated these needs with explicit attribute values. Users, who were choosing a laptop computer, were not asked to express preferences regarding explicit attributes (e.g. weight, storage size, processing power), but rather what they needed from a laptop computer (e.g. portability, data storage, gaming possibilities). Using pre-programmed associations, the system then knew which attributes to vary for every kind of customer need.

In a similar fashion, Hu and Pu (2009) used a personality questionnaire, rather than item ratings, to establish a user profile in a movie recommender system. It can be noted that this is a more abstract way of inferring preferences. They found that, while the perceived accuracy of the personality-based interaction method was not significantly better than that of the rating-based method, it did score better on user effort. Both the perceived user effort and the completion time of using the system were significantly lower for the personality-based recommender system. The study by Hu and Pu (2009) serves as an example confirming that people generally experience less effort disclosing context-independent information about themselves, compared to indicating specific attribute preferences within a decision domain.

Critiquing

A recommender system that uses critiquing allows its users to fine-tune their user profile by giving feedback upon receiving a recommendation. When users evaluate the recommendation and give concrete feedback on what aspects are not to their satisfaction, the system can search for a more suitable item, while taking the specified criteria into consideration. Critiquing can be implemented quite naturally in a conversational recommender system, as one might instinctively want to specify, based on certain features, what they would prefer over the current recommendation (Nguyen & Ricci, 2018).

Pommeranz, Broekens, Wiggers, Brinkman, and Jonker (2012) argued that asking explicit questions to help classify the user on a few superficial characteristics, and then updating this profile along the way, most naturally fits human decision making. Allowing the user to refine their profile during their interaction with the system is important, as users' preferences may change as they learn more about the decision domain (Pu et al., 2012) and as they interact more with the recommender system

(Pommeranz et al., 2012). This is in line with a framework regarding constructive consumer choice, proposed by Bettman, Luce, and Payne (1998). This framework indicates that consumers generally do not have consistent and clearly defined preferences, but rather construct these preferences when faced with a decision task, based on, among many other factors, the context of making the decision and the experience they have with the decision domain. From this framework, it can be argued that, when requesting a recommendation multiple times during a theme park visit, users' preferences for attractions may very well change as the day progresses. Offering support for critiquing has also been found to increase recommendation accuracy and user experience (Bostandjiev, O'Donovan, & Höllerer, 2012).

A trade-off between personalization and usability

Controllability versus complexity

Literature concerning the controllability of recommender systems describes a trade-off between controllability and complexity. All algorithmic approaches to calculate utility scores for recommendations rely on its users to disclose a certain level of information, to best mitigate the cold start problem (Felferning & Burke, 2008). In a literature overview, Jugovac and Jannach (2017) mentioned that the amount of control the user is given over the way their recommendations are computed positively affects their satisfaction with the produced recommendations. Additionally, the level of control positively affects the perceived utility of a recommender system, as stated by Bostandjiev et al. (2012). In their research, the level of control was defined as the extent to which the user was able to fine-tune the way their preferences were processed by the recommender system. However, the level of control that the user is offered is also positively associated with the complexity of the system (Jugovac & Jannach, 2017). It can be concluded that more elaborate and precise ways of indicating preferences can lead to better recommendations, at the cost of ease of use.

The elaborateness of the initial set of questions

The combination of an initial set of questions and a critiquing component may pose a similar trade-off. A more elaborate set of questions up front will better mitigate the cold start problem (Felferning & Burke, 2008), and thus produce more accurate recommendations from the start. However, it may come at the cost of a higher cognitive load. Potential users of a recommender system might not bother with it if it takes too long to indicate their preferences. Gathering less information up front, and implementing functionality for critiquing, may make it more attractive to start using the system while still providing personalized recommendations. However, personalization through critiquing is based on users' past responses to recommendations. The more concise set of questions up front may, therefore, initially

produce recommendations that are too generic and do not fit the specific user. This may lead to disappointment for a user that expects more personalized recommendations (Bobadilla et al., 2012).

The effects of domain knowledge

As concluded in a recent survey on conversational recommender systems (Jannach, Manzoor, Cai, & Chen, 2020), a clear understanding of what factors influence the helpfulness of a conversational recommender system is currently lacking. It is, for this reason, impossible to objectively point to an optimal balance in the trade-offs mentioned in the previous section.

A factor that may affect the optimal balance between controllability and complexity is the amount of knowledge that the user of the recommender system has about the decision domain and the item space. Payne, Bettman, Schkade, Schwartz, and Gregory (1999) described that in many choice situations, people only seem to exhibit a notion of well-articulated preferences when they are familiar with the relevant items of choice. This suggests that, while asking domain-specific preference questions may be useful (to some extent) for experts, it makes less sense for visitors who are not familiar with the items. This notion is in line with the research of Pommeranz et al. (2012), who found that participants are willing to indicate preferences in more detail when they are familiar with the items in question. Similarly, people who are not familiar with the decision domain find it challenging to state explicit preferences on an attribute-level and may be more comfortable indicating their needs in a more abstract manner (Randall et al., 2007).

Knijnenburg, Willemsen, and Broeders (2014) also proposed that different users may have different needs when interacting with a recommender system. In their paper, summarizing the results of four experiments, several different PE methods have been investigated in a recommender system proposing energy-saving measures. They found that expert users prefer systems where they can exert direct control over the attribute weights (e.g. attribute-based PE). This is hypothesized to be caused by the fact that these methods allow the user to better apply their intimate domain knowledge. PE methods that were found to be more suited to novice users offer less direct control to the user and instead rely on more indirect ways of gathering user preferences (e.g. needs-based PE). The reason for this may be that novice users have difficulty expressing their preferences explicitly. Randall et al. (2007) also found that novice users perceive needs-based PE as providing more comfort and as easier to use than attribute-based PE, whereas the opposite was found for expert users.

For Efteling, it is expected that different visitors may use a recommender system to satisfy different needs, depending on their familiarity with Efteling and the attractions that it has to offer. Based on the findings stated above, the optimal choice regarding the trade-offs mentioned earlier in this report may

differ among users that have different levels of domain knowledge. Someone who rarely visits Efteling may not feel knowledgeable enough to express specific preferences up front regarding the attractions in the park. Such a user may require a recommender system acting as an expert that shows them around. When it does, they may want to evaluate the recommendations using critiquing, as their preferences develop. An Efteling veteran, however, may know exactly what they want and will use the recommender as an assistant, which they want to configure precisely as they please. For this purpose, they may want to specify a lot of information while initializing their profile, to immediately receive worthwhile attraction recommendations. Critiquing is also expected to be useful for experts, but not as prominently, as their preference may not vary as much during their visit.

Research question and hypotheses

As described by He et al. (2016), little research has been conducted investigating whether different users might require different levels of control while interacting with a recommender system. This thesis describes a study that contributes to the research domain by investigating users of a theme park recommender system. Specifically, the study investigates whether users with different levels of domain knowledge require different levels of control while interacting with a recommender system. To evaluate this, the effects of controllability on users' self-reported level of satisfaction with the provided recommendations are investigated. Additionally, the effects on the persuasiveness of the recommender system are evaluated, in terms of the probability of recommendations being accepted. As such, the following research question is formulated:

Research question: "In what ways do controllability and domain knowledge affect user satisfaction and persuasiveness of a conversational recommender system for theme park visitors?"

Summarizing the earlier research described above, several hypotheses are established. In general, users who feel more in control over indicating their preferences are more satisfied with the produced recommendations, as described by Jugovac and Jannach, 2017. However, they also state that increased controllability can lead to more complexity in the system. Further, Knijnenburg et al. (2014) described how users with a high level of domain knowledge prefer interaction methods that offer a high level of control, and vice versa. Therefore, the following hypotheses are described:

Hypothesis 1a: The level of control that users perceive to be given over indicating their preferences positively affects recommendation satisfaction and persuasiveness of a recommender system.

Hypothesis 1b: Domain knowledge moderates the effects described in the first hypothesis, such that they are more positive for users with a higher level of domain knowledge.

Comparing several PE methods, Knijnenburg et al. (2014) found that the domain knowledge of users affects how comfortable they are indicating domain-specific preferences. It is expected that PE methods collecting more specific preferences will lead to more suitable recommendations for those who are able to disclose them. However, users who lack the knowledge to comfortably indicate these preferences will not be able to profit from this. This is reflected in the second hypothesis of this thesis:

Hypothesis 2: Users with a high level of domain knowledge prefer a PE method that allows them to directly apply specific knowledge about the decision domain. On the other hand, users with a low level of domain knowledge prefer a PE method that requires less domain-specific knowledge.

In a recommender system containing functionality for critiquing, a more elaborate set of questions up-front is expected to result in more suitable recommendations when users start to use a recommender system (Felferning & Burke, 2008) but may increase the cognitive load (Jugovac and Jannach, 2017). While the increased level of control is expected to have positive effects on satisfaction and persuasiveness in general, this effect may be less strong for users with a low level of domain knowledge, as they may rely more on the critiquing component to fine-tune their preferences as they are formed. Therefore, similar effects as those described for the perceived level of control in hypotheses 1a and 1b are expected to be found for the elaborateness of the initial set of questions:

Hypothesis 3a: The elaborateness of the initial set of questions positively affects recommendation satisfaction and persuasiveness of a recommender system.

Hypothesis 3b: Domain knowledge moderates the effects described in hypothesis 3a, such that they are more positive for users with a higher level of domain knowledge.

This thesis describes two experiments that aim to investigate the formulated hypotheses. Both experiments consist of manipulating aspects of the recommender system, and measuring the effects on perceived control, recommendation satisfaction, and persuasiveness for each user.

In experiment 1, two different PE methods, requiring different levels of domain-specific knowledge, are compared. The effects of the manipulation are evaluated to investigate hypotheses 1a, 1b, and 2. In the second experiment, the recommender system includes a critiquing component. In addition to

comparing different PE methods, the elaborateness of the initial set of questions is manipulated. All formulated hypotheses are investigated using the manipulations in the second experiment. An in-depth description of the manipulations and the set-up of the experiments are provided below.

Method

User-centric evaluation framework

In order to answer the research question that has been posed in this thesis, the user-centric evaluation framework, described in detail by Knijnenburg et al. (2012), is used. This framework was developed to provide a systematic approach to recommender system research. It is used to analyze how different aspects of interaction with a recommender system affect each other. The framework consists of six components, a brief summary of which is provided below. A visualization of the components, and how they relate to each other, is provided in Figure 1.

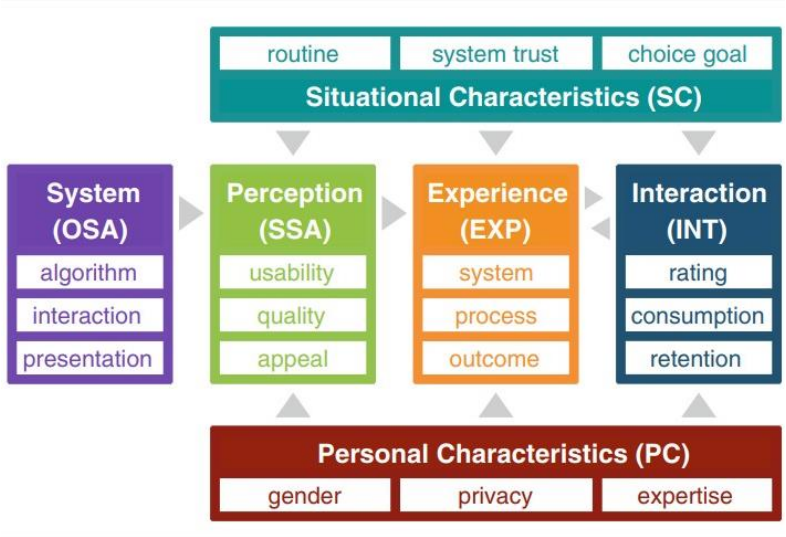

Figure 1: The user-centric evaluation framework (Knijnenburg & Willemsen, 2015)

Objective System Aspects (OSA) denote attributes of the system under investigation. These are elements of the recommender system that a developer can change to influence the user experience. OSAs include, among many other aspects, the appearance of the recommender system, the algorithm that is used, and the speed at which it responds to user input. OSAs directly influence *Subjective System Aspects (SSA)*. This component represents attributes of the system as they are perceived by the user. Inherently, they are different for each user of the recommender system. Examples of SSAs are aesthetic value, ease of use, and perceived speed. These aspects are important to assess whether a change in an OSA is perceived by users. SSAs are considered mediating variables between OSAs and both *User Experience (EXP)* and *Interaction (INT)*. EXPs are self-relevant evaluations of the recommendation system. These

can relate to different aspects of system usage. Examples of EXP variables are evaluations of the system itself, the process of using the system, or the recommended items. INT denotes the measurable aspects of users' interaction with the system. Examples include clicking behavior, time spent browsing, and acceptance rate of recommendations. These aspects are analyzed to research whether perceived differences also translate to differences in the evaluation of the system and to observable behavior. Experience and interaction aspects are regarded to be closely related to each other: a positive experience with the system can lead to an increase in interaction while, conversely, more interaction with the system may lead to a more positive experience. Along with the aforementioned components, *Personal and Situational Characteristics (PC and SC)* are considered in the model. These are characteristics of the user that are not influenced by the model but can affect the outcome of a manipulation. Examples of PCs include the personality or demographic information of the user, whereas examples of SCs are the level of trust in the system or the goal that the user has while using the system.

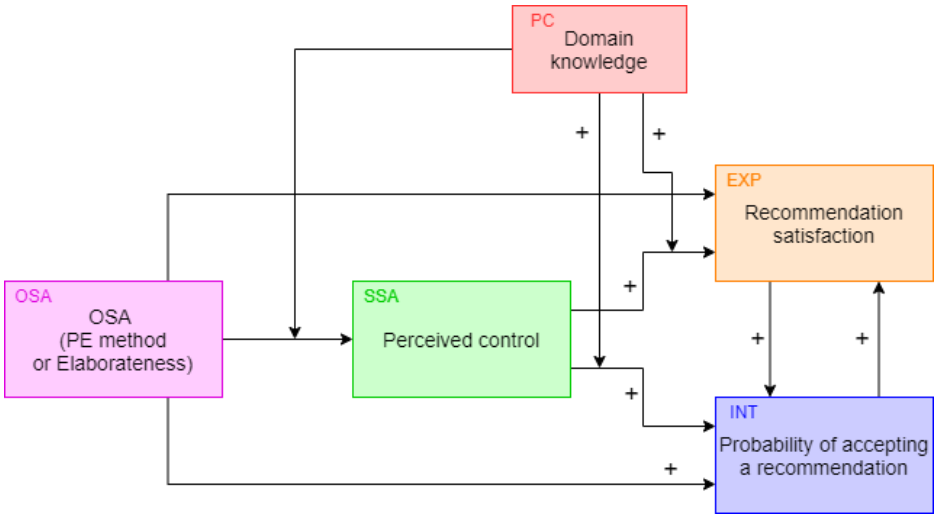

Figure 2: The current research model.

As the current thesis employs the user-centric evaluation framework, the relevant components of the research question and accompanying hypotheses can be embedded in the framework. The resulting model with the expected relations are displayed in Figure 2. Two OSAs are evaluated in hypotheses 2 and 3a: the PE method used and the elaborateness of the initial questionnaire. The effects of these OSAs on both EXP (in terms of recommendation satisfaction) and INT (in terms of the probability of a recommendation being accepted) are investigated in this thesis. Furthermore, perceived control is the amount of control users feel they are given while indicating their preferences. As such, it is an SSA. As described by hypothesis 1a, perceived control is expected to affect users' EXP and INT. Furthermore, it

could serve as a mediator for the effects of the identified OSAs on EXP and INT, which the model also assesses. Lastly, domain knowledge is a PC that is expected to serve as a moderator of the effects described above, as described in hypotheses 1b, 2, and 3b.

Materials

Recommendation engine

The recommender system that was developed by Efteling and the recommendations it provides are based on an underlying recommendation engine that is the product of a collaborative project between Efteling and Jheronimus Academy of Data Science (Abbaspourghomi, 2020; Orense, Chandrasekaran, Yolande, & Rashipour, 2020). As part of this project, participating visitors used an app that tracked their movements throughout their day and included a survey to extract user features and ratings. The research was conducted with the goal of understanding the behavior of visitors and provided the basis of several insights that the engine relies on. Additionally, the results of Orense et al. (2020) were used to determine the optimal user features for the recommender system to gather while establishing a user profile.

The resulting underlying model that is employed by the recommendation engine incorporates a number of features that are deemed important to calculate the optimal recommendation at any given time. The model incorporates a hybrid approach. This means that it considers both user features, such as the user's group composition and a set of attractions that the user likes, as well as contextually relevant attraction features, such as the current waiting time, crowdedness, and walking distance to the attraction. A ranking algorithm generates a ranked list of all attractions based on the included features. The highest-ranking attractions are sent to the recommender system upon request, such that these can be recommended to the user.

Recommender system

Deployment

The recommender system that is developed by Efteling is called 'Pennenveer', which is also the name of the digital assistant that Efteling uses on their website. Both experiments described in this thesis are conducted using Pennenveer. It was launched as part of a smartphone application called Efteling Lab: an initiative from Efteling to enable interested visitors to try out new digital products that are not yet finished and to receive feedback. Efteling Lab was initially introduced for the sole purpose of distributing Pennenveer. As such, Pennenveer was the only service included in Efteling Lab upon its release. Eventually, Efteling intends to incorporate Pennenveer into the main Efteling app. The product was

launched for iOS devices on the 20th of October 2020. During the time frame of the study presented in this thesis, the only supported language of the application was Dutch.

Functionality

Pennenveer is a smartphone application that aims to assist users to get more out of their visit to Efteling by providing contextually relevant recommendations for which attraction to visit next. For groups of people visiting Efteling, the app only needs to be installed on one device to receive recommendations that are tailored to the group as a whole.

The application features a conversational recommender system, in which users can choose between a set of pre-determined answer options to converse with the system. After installing the application and opening it, the user is prompted to fill out an initial questionnaire to establish a personal user profile. The initial questionnaire contains the manipulations of the PE method and elaborateness of the questionnaire, which are used to investigate the hypotheses formulated in this thesis. A detailed description of the manipulations is provided in the subsection 'Experiments'. In addition to the manipulations, the subjective system aspect 'perceived control' and the personal characteristic 'domain knowledge' are also measured in the initial questionnaire as questionnaire items. The full contents of the initial questionnaire, including the phrasing of the questions, are displayed in Appendix A. Screenshots of the user interface of the application and how the questions are displayed are provided in Appendix B. After the user completes the initial questionnaire, the first recommendation is given, provided that the user is currently in the park. At any point during their visit, the user can request a recommendation from Pennenveer.

Figure 3: A recommendation.

A recommendation is provided in the form of a suggestion for a single attraction, by showing its name and picture, a description of the type of attraction, and the current waiting time. An example of a recommendation is shown in Figure 3. Upon receiving a recommendation, a user can explicitly indicate whether they accept the recommendation or not. Acceptance prompts the application to launch the main Efteling app, allowing the user to immediately navigate to the recommended attraction. When a recommendation is rejected, the recommender system enables the user to either ask for a new recommendation immediately or to pause the conversation for the time being. If the latter is chosen, the recommender system allows the user to request a new recommendation at their convenience. Screenshots of the way recommendations are provided in the application are provided in Appendix B.

To measure users' recommendation satisfaction, users are requested to fill in a final questionnaire at the end of their visit. This questionnaire can be triggered by the user by indicating that they are leaving the park. Additionally, starting one hour before the park closes, an hourly push notification is sent up to three times if the user has not filled in the questionnaire yet, asking them whether they have time to fill in the questionnaire. The final questionnaire is also used to collect feedback about users' experiences, both regarding the recommender system and their visit to the park in general. In Appendix A, the full set of questions included in the final questionnaire, as they are phrased in Pennenveer, is displayed. Screenshots of how the final questionnaire looks are provided in Appendix B.

Experiments

To verify the validity of the hypotheses stated in this thesis, two aspects of the recommender system are manipulated in a way that is expected to affect the controllability of the system. The first experiment compares two different PE methods. In the second experiment, a functionality for critiquing is added to Pennenveer. Both different PE methods and a varying elaborateness of the initial set of questions are investigated in experiment 2. An in-depth description of both experiments is provided below.

Experiment 1

As mentioned in the subsection 'Recommendation engine', one of the user features that the engine requires to provide personalized recommendations is a set of attractions that the user and their group enjoy. These attraction preferences are gathered as part of the initial questionnaire using a PE method. The first experiment compares two PE methods that require different levels of domain-specific knowledge: exemplar-based and needs-based (a description of each method is provided below). Both of these methods result in a set of up to five attractions, which is sent to the recommendation engine. Using A/B testing, users are randomly assigned to one of the two experimental groups, both with a probability of 50%. In Figure 4, the user interface of both preference elicitation methods is displayed.

Exemplar-based preference elicitation

The first PE method that is implemented collects users' preferences by allowing them to choose between a set of exemplars. Users with this PE method, displayed in Figure 4, are shown a list of nine specific attractions that can be found in Efteling, of which they can select up to five to indicate that they like them. It is reasoned that this form of preference elicitation offers a relatively high level of control to those who are familiar with the attractions on display, as users can indicate for specific attractions that they like them. This could be particularly useful in the theme park domain, where users may have certain favorites that they would not like to miss during their day. However, it does require knowledge of the various attractions in Efteling. Therefore, the PE method is expected to be most suitable for users

with a high level of domain knowledge, as unfamiliar visitors might be unable to answer this question, or they might select less than five attractions and thus send less detailed user information to the recommendation engine.

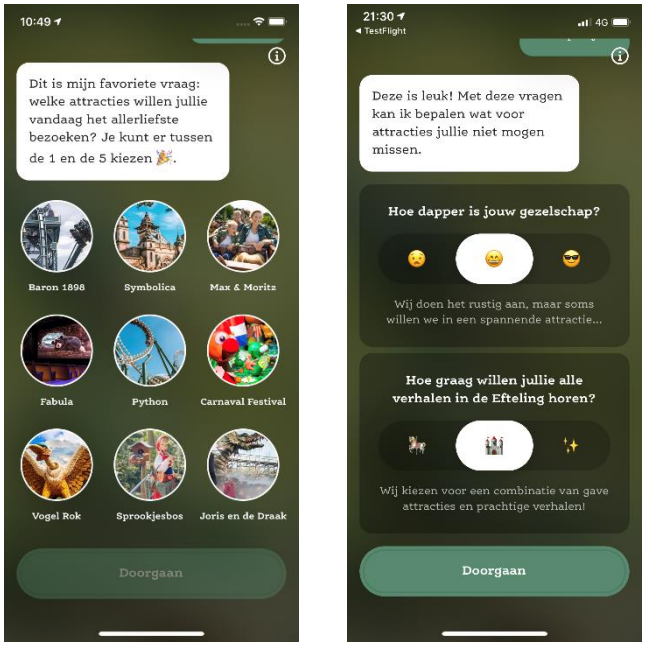

Figure 4: Exemplar-based (left) and needs-based (right) attraction preference elicitation methods.

Needs-based preference elicitation

The second group of users included in the first experiment receives a different method of indicating their preferences. These users are prompted to answer two questions, displayed in Figure 4, about the needs of the group regarding their visit. The first question is used to measure the user's need for thrill-seeking attractions, whereas the second question is used to inquire about the user's need for storytelling elements in the attractions. As the recommendation engine requires a set of specific attractions as a user feature, the answers to these questions have to be converted to a set of attractions. Both questions are answered by selecting one of three options, resulting in nine different possible configurations of answers. Using expert knowledge of employees of Efteling, each configuration is in the background linked to a set of five specific attractions that are balanced to reflect the indicated user needs. These five attractions are sent as attraction preferences to the recommendation engine. The set of attractions that are associated with each configuration of answers are listed in Appendix C.

The needs-based PE method offers less direct control to the user, as users can only answer two questions about themselves, leaving less room for personalized input. The method, however, does not

require domain-specific information. As such, it is expected to be more suitable for visitors with a low level of domain knowledge. Additionally, the needs-based PE method always sends five attractions to the recommendation engine on the user's behalf. This is in contrast with the exemplar-based PE method, which allows choosing a set of fewer than five attractions. By using expert domain knowledge of Efteling employees, it is expected that these five attractions reflect the needs of the user, even if that user does not know these attractions.

Critiquing component

The second experiment featured in this study contains a critiquing component that allows users to fine-tune their profile during their visit. Before elaborating on the manipulations included in experiment 2, this subsection explains how the critiquing component was implemented in the recommender system.

Recommendation engine weights

Part of the functionality of the ranking algorithm in the recommendation engine is making a trade-off between various attributes that may be important to the user. An example of this is deciding between recommending a nearby attraction with a long waiting line, or an attraction that has no line but is quite far away. Both waiting time and walking distance are contextual features that can affect the suitability of an attraction (e.g. Li, 2010; Lathia, 2015). The recommendation engine features a number of weights to indicate the importance of certain contextual attraction features. These features are calculated for each attraction upon requesting a recommendation. The weights, having a default starting value of 1, can be changed per user. The critiquing component that is added in experiment 2 manipulates the weights for specific users. By doing so, it personalizes the relative importance of the associated features. Table 1 displays the features that are associated with an individually manipulated weight.

Feature	Explanation
Walking distance	The walking time in minutes.
Waiting time	The waiting time of the attraction in minutes.
Probability	The algorithmically predicted probability of the attraction being visited next, based on various factors (e.g. user preferences, group composition, and the attractions that were already visited).

Table 1: The features in the recommendation engine that are associated with a weight that is manipulated in the critiquing component.

Critiquing

When users reject a recommendation provided by Pennenveer, they are asked to provide feedback on their reason for rejecting it. In the first experiment, this feedback is not automatically processed to

personalize the recommendations for that particular user. The user is made aware of this after providing the feedback. In experiment 2, a critiquing component processes this feedback to further personalize the user profile. As described in Table 1, the recommendation engine that provides the recommendations features three weights: a weight for walking distance, a weight for waiting time, and a weight for probability. When a user indicated that they found the recommended attraction to be too far away from their current location, the recommender system made the generalizing assumption that the user found walking distance to be an important aspect of a recommendation. When this happened, the recommendation engine would increase the weight of walking distance for that user by 0.2 (the default starting value is 1 for each weight). The value of 0.2 was chosen with the intent to subtly optimize the importance of the features for each user over time, without drastically changing the nature of the recommendations with each critique.[1] A similar assumption and weight change were made when a user indicated that the waiting time for an attraction was too long. When a user rejected a recommendation on the basis of not liking the attraction, the assumption was made that the recommender system relied too heavily on waiting time or walking distance and increased the weight of probability by 0.2 to compensate for this. Other feedback options that the user could choose ('I already visited the attraction', 'Not now, maybe later', and 'Other') did not affect any weight in the recommendation engine.

Experiment 2

The implementation of the critiquing component not only allows users to fine-tune their profile during their visit, but it also adds the importance of short waiting times and walking distances as a user feature that can be personalized. The second experiment features a 2x2 design, in which two manipulations are conducted simultaneously. The first manipulation is the same as in experiment 1: half of the users interact with an exemplar-based PE method, whereas the other half receives a needs-based PE method. The second manipulation involves varying the elaborateness of the initial questionnaire.

The elaborateness of the initial questionnaire

In the recommendation engine, the default starting weights for preferences of waiting time and walking distance are set to 1 for each user. In the second experiment, 50% of users, chosen at random, can change these values up front by answering two more questions in the initial questionnaire. By use of these questions, displayed in Figure 5, users can indicate the importance of receiving recommendations for attractions that have relatively short lines, and for attractions that are nearby. Similar to the needs-based PE method for attraction preferences, the importance of each feature can be indicated by

[1] Due to time constraints imposed on the presented study, the optimal values for altering the weights have not been empirically validated.

choosing one of three answer options. By indicating that a certain feature is particularly unimportant, the starting weight of that feature is set to 0.5. If the feature is particularly important, the starting weight is set to 1.5. When the neutral option in the middle is chosen, the starting weight remains 1. These values have been chosen to take into account user preferences from the start, which users can then fine-tune during their interaction with the system.[2]

By giving users the freedom to manipulate the importance of the mentioned user features up front, they are offered more control in initializing their profile. It is expected that this results in more personal, and therefore useful, recommendations when the user starts using Pennenveer. However, the feature does add two questions to the initial questionnaire, requiring more effort from the user before being able to start using the system. It is expected that the effects of a more elaborate questionnaire are more positive for users with a high level of domain knowledge, as users with a low level of domain knowledge may have less well-articulated preferences prior to their visit (Payne et al., 1999).

Figure 5: Questions to indicate preferences for waiting time and walking distance.

Measures

Several constructs are measured during users' interaction with the recommender system to investigate the research model posed in this thesis. These constructs, and their expected relations, are also illustrated in Figure 2.[3] Below, these constructs are listed and their measurements are explained. The questionnaire items have been created in a way that fits the tone that is used in Pennenveer, and that is believed to best measure the construct under investigation. The Dutch phrasing of the questionnaire items, as they are included in Pennenveer, are displayed in Appendix A.

Perceived control (SSA)

This subjective system aspect denotes the extent to which the user found they were in control of indicating their preferences about the attractions that they would like to visit. It is an inherent aspect of the recommender system, as perceived by the user. This construct is measured as an item in the initial questionnaire, positioned directly after the attraction preference elicitation method and, if applicable, after the additional question concerning users' preferences of waiting time and walking distance.

[2] Due to time constraints imposed on the presented study, the optimal values for altering the weights have not been empirically validated.

[3] It should be noted that the constructs are measured using single-item questions, whereas it is advised to include multiple items per measured construct (Knijnenburg et al., 2012). However, as the research was conducted on a live product, this was deemed infeasible.

Because the question is part of the initial questionnaire, the measure of perceived control contains no information about the potential effects of the critiquing component that is added in experiment 2.

An English translation of the phrasing of the questionnaire item is as follows: “One question left: How much control do you feel you had over indicating which attractions you would like to visit?” The question is answered on a five-point scale, ranging from “Very little” to “Very much”.

Recommendation satisfaction (EXP)

This construct denotes the experience that the user has with the recommender system. Specifically, it measures the overall satisfaction that the user experienced with the recommendations that have been provided to them during their visit. This question is a part of the final questionnaire, and a translation of the phrasing is as follows: “How satisfied are you with the recommendations for attractions that Pennenveer has given you?” The question is answered using a star rating, in which users can give 1 to 5 stars to indicate their satisfaction.

Probability of accepting a recommendation and acceptance ratio (INT)

This construct is related to the persuasiveness of the recommender system. For each recommendation that the recommender system gives to a user, the user is asked whether they would like to go to the recommended attraction or whether they would like to visit a different attraction. A binary value is stored for each recommendation, indicating whether the recommendation has been accepted or not. Additionally, an acceptance ratio is calculated for each user, reflecting the number of recommendations that they accepted during their visit, divided by the total number of recommendations that they received.

Domain knowledge (PC)

This personal characteristic indicates how knowledgeable a user is about Efteling and the attractions that the park has to offer. To approximate this construct, a proxy is measured that Efteling uses to differentiate different segments of their visitors: visiting frequency. Visiting frequency is measured in the initial questionnaire, and a translation of the phrasing is as follows: “How often do you visit Efteling, on average?” This question can be answered with the following intervals: “This is my first visit to Efteling”, “Once per six months or more frequently”, “Once per year”, “Once per 2 years”, and “Once per 3 years or less frequently”.

Data collection

The public release of Pennenveer took place on the 20th of October 2020. On the 26th of October, the first A/B-test started, in which two alternative PE methods were compared. At first, the measurements

for perceived control and visiting frequency were collected as part of the final questionnaire to minimize the required effort in the initial questionnaire. However, these items were soon moved to the initial questionnaire. The first reason for this is that many users were found to not fill in the final questionnaire. This can be explained by the fact that the recommender system had no practical use for visitors at the end of their visit, potentially causing users to forget about the application or to not be willing to provide feedback. For example, of all users who received at least one recommendation, 74.9% did not disclose their perceived level of control. The second reason for moving the items to the initial questionnaire is that the validity of the measure for perceived control was deemed questionable. Users may have forgotten their exact perceptions during the initial questionnaire when a question about it is asked at the end of their visit. Shortly after releasing Pennenveer, Efteling was forced to close for two weeks, because of measures related to the COVID-19 pandemic. After this period, on the 25th of November, the items that measure perceived control and visiting frequency were moved to the initial questionnaire. Because of the reasons described above, data that was collected before the 25th of November was not included in the analysis.

The critiquing component, as well as the manipulation of the elaborateness of the questionnaire, was deployed to the recommender system on the 3rd of December at 11:15 a.m. This moment marks the end of experiment 1 and the start of experiment 2. Data collection for experiment 2 continued until the 14th of December, when Efteling was again forced to close due to measures related to the COVID-19 pandemic. This pandemic had multiple effects on the presented study that should be taken into consideration when interpreting the results, a full elaboration of which is discussed in the limitations section of this thesis.

Experiment	Start date	End date	*n* (full sample)	*n* (visiting frequency)	*n* (perceived control)	*n* (satisfaction)
1	25-11-2020	03-12-2020 (11:15 a.m.)	219	206	207	58
2	03-12-2020 (11:15 a.m.)	14-12-2020	258	258	258	63

Table 2: Summary of gathered data from experiments 1 and 2, including the sample size for each experiment and the number of collected observations for each questionnaire item.

After collecting the data, preprocessing was conducted on the dataset. The resulting datasets only contain data from users who at least received and responded to one recommendation, such that only

users are included who have visited the park and interacted with the recommender system. Data that was generated by developers of the recommender system was removed. This data was marked by the developers in a feedback comment in the questionnaire. Lastly, for each user, the acceptance ratio was calculated by dividing the number of accepted recommendations by the total number of received recommendations. A summary of the resulting datasets of the questionnaire data, including the timeframes and the number of records for each measured construct is displayed in Table 2. For data regarding the provided recommendations and whether they were accepted, 806 records were included in the analysis for experiment 1, and 805 for experiment 2. The data was analyzed using R (version 4.0.2). All tests are conducted with a significance level of $\alpha = 0.05$.

Results

Data description

Before evaluating the hypotheses presented in this thesis, a description of the collected data is provided along with a comparison between the data gathered in each experiment.

Domain knowledge

Visiting frequency, as a proxy for domain knowledge, was measured on a five-point scale. Table 3 displays the distribution of observations for each answer option for each experiment.

Answer option	Experiment 1	Experiment 2
Every six months or more frequently	50	61
Every year	86	108
Every two years	29	34
Every three years or less frequently	33	32
This is my first visit	8	23

Table 3: The frequency of observations for each value of visiting frequency, in both experiments.

Based on visiting frequency, a user's domain knowledge is included in the analysis as a dichotomous variable. The variable reflects whether a user is an expert (who visits Efteling every year or more frequently), or a novice (who visits Efteling every two years or less frequently). This distinction between the two values provides the most even split of data points into each condition. There were approximately twice as many experts as there were novices in the data, with 66.5% of users in experiment 1, and 65.5% of users in experiment 2. The proportion of experts in each experiment is considered equal, $\chi^2(1) = 0.014$, $p = 0.91$.

Perceived control

The subjective system aspect of perceived control is measured using a five-point scale. Figure 6 displays the distribution of the collected values for perceived control in each experiment. By inspection of the plots, and by applying the Central Limit Theorem, the data is assumed to be normally distributed. As an F-test to compare two variances indicates that the two distributions are of equal variance, $F(206, 257) = 0.96$, $p = 0.78$, a Student's t-test may be used to compare the means of the distributions. This analysis reveals that the means in experiment 1 ($M = 3.222$, $SE = 0.064$) and experiment 2 ($M = 3.217$, $SE = 0.059$) can be considered equal, $t(463) = 0.059$, $p = 0.95$, $d = 0.006$.

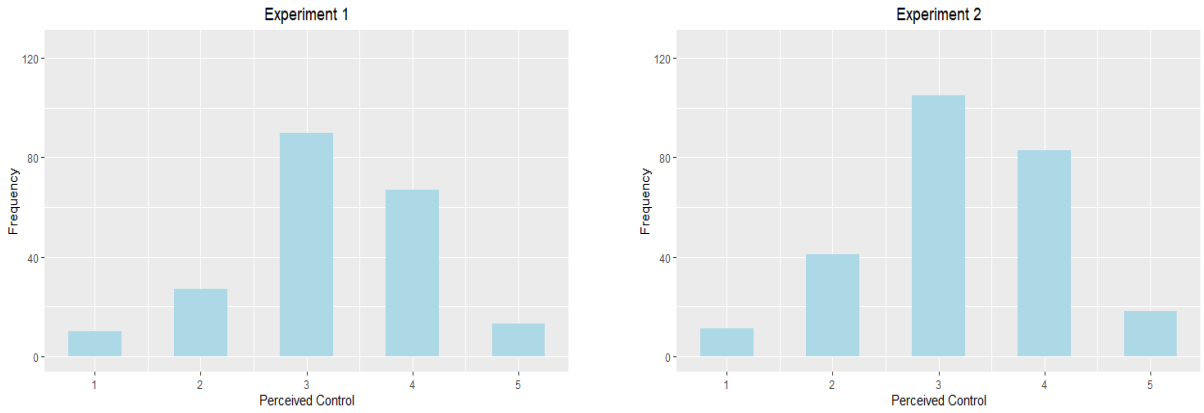

Figure 6: Distribution of perceived control in both experiments.

Recommendation satisfaction

This experience variable is measured on a five-point scale, as part of the final questionnaire. Figure 7 displays the distribution of recommendation satisfaction in each experiment. An F-test to compare the variances reveals that the variances of recommendation satisfaction among the experiments may be considered equal, $F(57, 62) = 1.07$, $p = 0.79$. A Student's t-test is used to compare the means of the two experiments, as it is regarded to be robust to deviations from normality by application of the Central Limit Theorem. Recommendation satisfaction that was measured in experiment 1 ($M = 2.29$, $SE = 0.17$) was found to be significantly lower than that measured in experiment 2 ($M = 2.76$, $SE = 0.16$), $t(119) = -2.04$, $p = 0.043$, $d = 0.37$. However, this result is not supported by a non-parametric Wilcoxon-Mann-Whitney test, which did not indicate values from experiment 1 to be significantly different from values in experiment 2, $W = 1598.5$, $p = 0.10$.

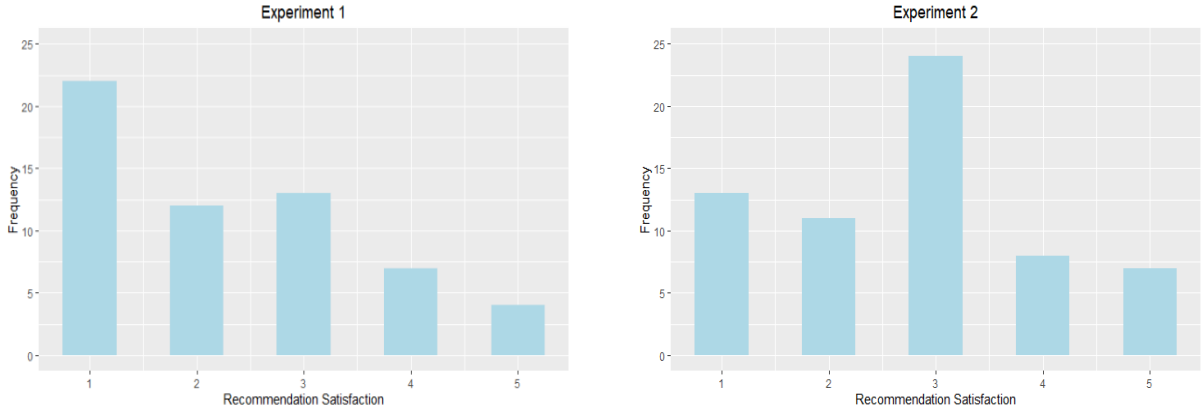

Figure 7: Distribution of recommendation satisfaction in both experiments.

Ratio of acceptance

To visualize the recommendation acceptance rate of users in both experiments, the ratio of acceptance is calculated. This variable denotes the number of recommendations users accepted, divided by the number of recommendations they received. The distribution of this variable for each experiment is displayed in Figure 8. Notably, a large proportion of users did not accept any recommendations, in both experiment 1 (60.7%) and experiment 2 (45.3%). Note that only users who received and responded to at least one recommendation were included in the analysis. The values for acceptance ratio in experiment 1 (*Mdn* = 0) were found to be lower than in experiment 2 (*Mdn* = 0.2), using a Wilcoxon-Mann-Whitney test, $W = 24334$, $p = 0.005$.

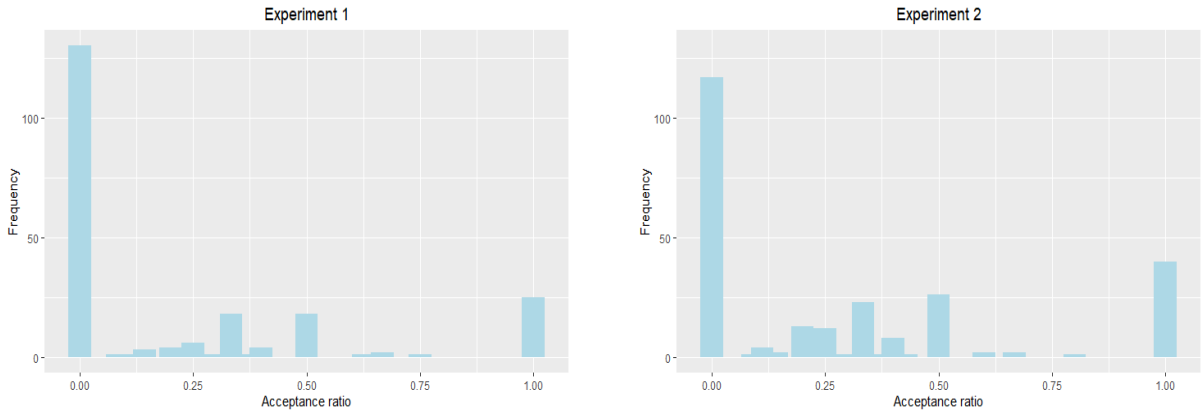

Figure 8: Distribution of acceptance ratio in both experiments.

Analysis

To investigate the research question and hypotheses posed in this thesis, the collected data is analyzed to research effects of controllability on users' experience and behavior in both experiments, as well as possible moderation by domain knowledge. As the effects on experience and behavior are investigated using different statistical tests, the effects on recommendation satisfaction are analyzed first, followed by the effects on the probability of accepting a recommendation.

First, visualizations of the data for various conditions are inspected to identify certain effects. To verify these findings, as well as to further investigate the data, a path model is created for both experiments in R using the Lavaan package, using bootstrapping (1000 replications) to obtain standard errors. Using this model, the hypothesized effects on recommendation satisfaction are analyzed. For both experiments, a model is fitted that includes the full sample under investigation. Then, stratified models are created for each level of domain knowledge. Using these, the specific effects for novices and experts

are revealed, which can be interpreted to assess moderation. To statistically assess moderation, Sobel tests are calculated to investigate whether the effects for novices and experts differ significantly (Cherrie, 2019). To inspect the full results, the regression tables of the (stratified) path models for each experiment can be inspected in Appendix D.

Lastly, the hypothesized effects on behavior, in the form of the probability of accepting a recommendation, are assessed using mixed logistic regressions, with device id as grouping factor. These regression models, along with visual plots of the data, are used to investigate whether the manipulations, as well as perceived control, are significant predictors for the probability that a recommendation is accepted. In addition, interaction effects of domain knowledge with each of the mentioned effects are analyzed to investigate moderation. Appendix D displays the regression tables for the mixed logistic regression in both experiments. Finally, a mixed logistic regression is conducted for each experiment to investigate whether the probability of accepting a recommendation is related to recommendation satisfaction, to assess the interplay between the two outcome variables.

Missing data

As displayed in Table 2, a large number of observations in the questionnaire data have missing values for recommendation satisfaction in both experiment 1 (73.5%) and experiment 2 (75.6%). To retain as much information in the analyses as possible, Full Information Maximum Likelihood (FIML) is used for the path models. FIML has been shown to perform better than, for example, listwise deletion (i.e. only including observations that contain no missing value for any of the relevant variables) on the efficiency of parameter estimates and type-1 error rates, especially under large proportions of missing data (Enders & Bandalos, 2001). FIML does assume that data is normally distributed, although negative effects of nonnormality were found to be mitigated when using bootstrapping to estimate standard errors (Enders, 2001). Another assumption for both FIML and listwise deletion is that the missing data is not nonignorable, meaning that the missingness of an observation is not caused by the value of the observation. FIML or listwise deletion may introduce different biases when analyzing nonignorable data. For this reason, the results of the path analyses under listwise deletion have been reported in Appendix E. Notable differences between the results of the two missing data methods are mentioned in the current section. The validity of the assumption of ignorable missing data, and the implications of potential violations, are discussed in the limitations section of the discussion.

Experiment 1

Effects on recommendation satisfaction

The data collected during the first experiment is used to evaluate hypotheses 1a, 1b, and 2. Hypotheses 1a and 1b reflect the expectations that perceived control positively affects recommendation satisfaction, and that this effect is more positive for experts than for novices. To visualize the collected data, a median split is performed on perceived control, and the mean level of recommendation satisfaction for each group and each PE method is displayed in a plot in Figure 9. Note that this plot contains only information about the subset of observations for which recommendation satisfaction is collected ($n = 55$). Comparing the relative levels of perceived control, users with a high level of perceived control in the sample were more satisfied with the recommendations provided by Pennenveer, than users with a low level of perceived control. This finding is confirmed by the path model of a mediation analysis, displayed at the top of Figure 11, showing that perceived control indeed has a positive effect on recommendation satisfaction, with an estimate of 0.39 ($SE = 0.19$, $p = 0.038$). The findings support Hypothesis 1a.

The stratified models, displayed at the bottom of Figure 11, reveal that perceived control does not significantly affect recommendation satisfaction for both novices and experts. To assess whether these effects, while non-significant, are statistically different from each other, a Sobel test is conducted. The results of the test, displayed in Table 4, reveal that these effects are also not statistically different ($p = 0.38$). Therefore, no support for hypothesis 1b is found.

Hypothesis 2 suggests that the exemplar-based PE method produces more suitable recommendations for expert users, whereas the needs-based PE method is more suitable for novice users. The plot displayed in Figure 9 does suggest that the exemplar-based PE method is the preferred method of eliciting attraction preferences, but only for users who perceive a relatively high level of control. This may be explained by the fact that the exemplar-based PE method requires knowledge about the attractions that are displayed. It was reasoned that the PE method is less useful for users who do not possess this knowledge, leading them to provide less personal information to the system. In Figure 10, the mean level of perceived control is displayed for users with different PE methods and different levels of domain knowledge. The levels of perceived control plotted are quite similar, except for novices who used the exemplar-based PE method, who perceived notably smaller levels of control. The plot is supported by the stratified path models at the bottom of Figure 11, where PE method is shown to have a negative effect on perceived control for novices, with an estimate of -0.53 ($SE = 0.19$, $p = 0.004$), while no such effect is found for experts, with an estimate of 0.024 ($SE = 0.16$, $p = 0.88$). These effects are statistically different from each other ($p = 0.020$), as revealed by the Sobel test displayed in Table 4.

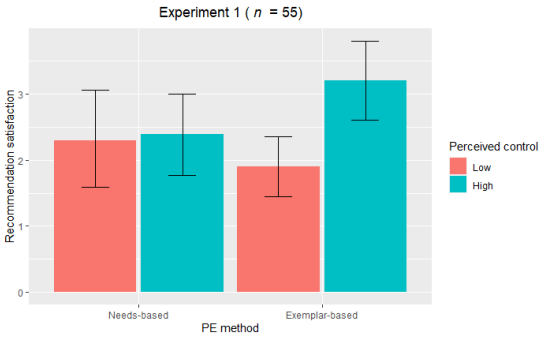

Figure 9: The mean level of recommendation satisfaction in experiment 1, displayed for each PE method and each relative level of perceived control, divided by a median split.

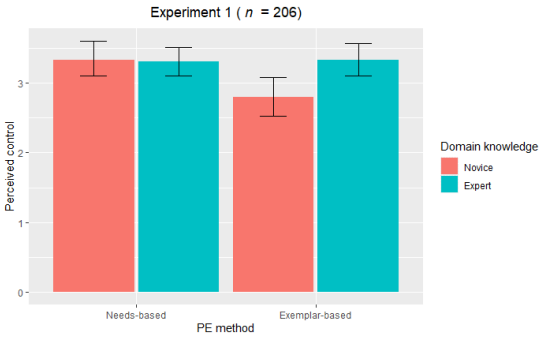

Figure 10: The mean level of perceived control in experiment 1, displayed for each PE method and each level of domain knowledge

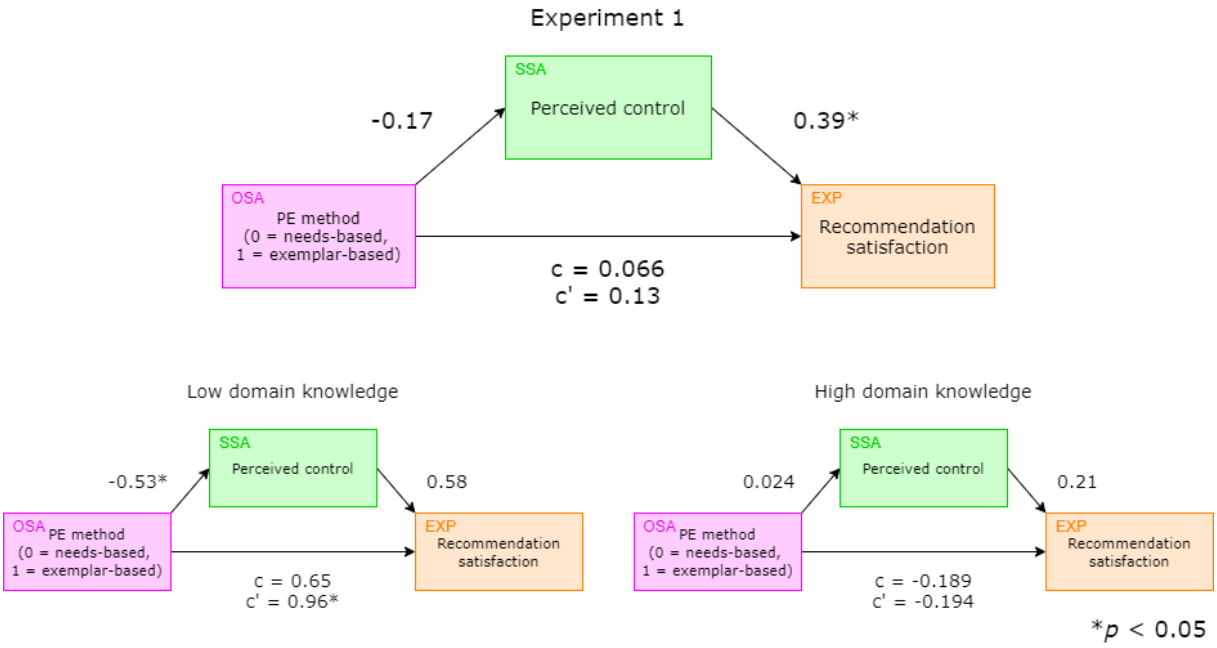

Figure 11: Results of a mediation analysis of experiment 1 (top), and the mediation analyses at each level of domain knowledge (bottom). For each effect, the regression estimate is displayed. An asterisk denotes a statistically significant effect. Appendix D displays the regression results in more detail.

However, the path model also reveals that for novices, a positive direct effect of PE method on recommendation satisfaction is shown, with an estimate of 0.96 (*SE* = 0.46, *p* = 0.037). An interpretation of the findings for novice users could be that, while novices generally prefer the exemplar-based PE method over the needs-based PE method, this effect is weakened by users who perceived a lack of control while using it, explaining the lack of a total effect of PE method on recommendation satisfaction for novice users. Interestingly, the stratified mediation model does not show significant effects for experts, indicating that both PE method and perceived control do not seem to significantly influence recommendation satisfaction. As revealed by the Sobel tests displayed in Table 4, neither the direct effects (*p* = 0.070) nor the total effects (*p* = 0.16) of PE method on recommendation satisfaction were significantly different between the levels of domain knowledge.

Hypothesis 2 is not supported by the data from experiment 1. For expert users, using either an exemplar-based PE method or a needs-based PE method does not affect recommendation satisfaction or perceived control. Novice users find the exemplar-based PE method to produce more satisfying recommendations, when controlling for perceived control. However, this effect is mitigated by novices who perceive less control, possibly due to a lack of domain knowledge.

Effect	Estimate	*SE*	*p*
Total effect PE method -> satisfaction	0.836	0.594	.160
Direct effect PE method -> satisfaction	1.152	0.635	.070
PE method -> perceived control	-0.557	0.239	.020*
Perceived control -> satisfaction	0.377	0.429	.379

Table 4: Results of Sobel tests to analyze differences of effects in experiment 1 between novices and experts. A positive estimate indicates that the effect is more positive for novices, compared to experts.

Lastly, similar (stratified) mediation models of the data from experiment 1 have been created using listwise deletion, instead of FIML, to estimate the effects under investigation. The results of these analyses are displayed in Appendix E. Using listwise deletion, a negative effect of PE method on perceived control was found for the full sample, with an estimate of -0.60 (*SE* = 0.22, *p* = 0.006). This implies that the subset of users who disclosed their recommendation satisfaction felt less in control while using the exemplar-based PE method, compared to the needs-based PE method. The total effect was still non-significant (*p* = 0.61). Apart from this finding, the results were qualitatively similar.

Effects on probability of acceptance

To inspect whether the manipulation in experiment 1 or the user's level of perceived control affects the probability that a recommendation is accepted, a mixed logistic regression analysis is conducted (n = 611), with device id as a grouping factor (k = 210). This analysis is conducted on the data of the recommendations that were provided, including whether they were accepted or not. Performing a logistic regression analysis on the provided recommendations, all recommendations can be analyzed while taking into account the fact that users are can respond to multiple recommendations. In Figure 12, the resulting effects are visualized in one model. The analysis revealed no effects of PE method or perceived control on the probability of a recommendation being accepted. Additionally, no moderation by domain knowledge was found. The model does not support the hypotheses that were investigated in experiment 1. In addition, a plot of the acceptance ratio of users, for each PE method and each level of domain knowledge is displayed in Figure 13. No conclusions can be drawn from inspecting the plot, due to the large variance in the data, as reflected by the error bars. A reason for the large variance may be the fact that a large proportion of users (60.7%) had an acceptance ratio of 0 in experiment 1.

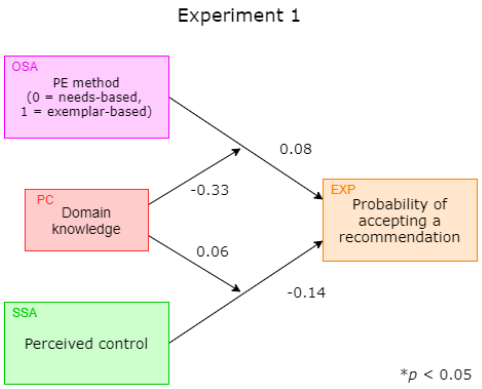

Figure 12: Logistic regression model with estimates of experiment 1. Detailed regression results are displayed in Appendix D.

Finally, a mixed logistic regression was conducted to assess whether there was a relation between recommendation satisfaction and the probability of acceptance in experiment 1. The resulting model (n = 319, k = 77) revealed that recommendation satisfaction was a significant predictor for probability of acceptance, with an estimate of 0.40 (SE = 0.16, p = 0.012). Note that, based on this finding, it cannot be concluded that recommendation satisfaction, as measured in the experiment, has a causal effect on acceptance rate, since recommendation satisfaction is measured as part of the final questionnaire, after

the recommendations have been received and responded to. Conversely, it may be argued that users' acceptance rate may have a positive effect on their reported recommendation satisfaction.

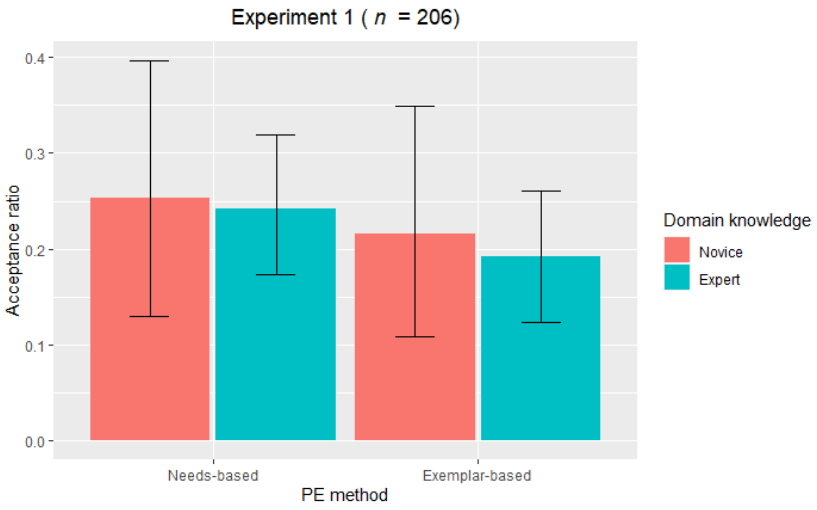

Figure 13: The mean acceptance ratio in experiment 1, displayed for each PE method and for each level of domain knowledge.

Experiment 2

Effects on recommendation satisfaction

Investigating the effects of the manipulations in experiment 2, all hypotheses formulated in this thesis can be evaluated. First, the results are analyzed to investigate whether perceived control positively affects recommendation satisfaction and whether this effect is more positive for expert users. This way, the validity of hypotheses 1a and 1b is assessed. A median split is performed on perceived control, and the mean level of recommendation satisfaction for each group is displayed in Figure 14, both for each level of PE method and each level of elaborateness. Again, these plots only include observations for which recommendation satisfaction is collected (n = 63). Both plots do not suggest that there is a significant difference between users with a relatively high level of perceived control and users who perceived relatively little control. The path model of the mediation analysis, displayed at the top of Figure 16, supports this finding (p = 0.62). This finding is not in support of hypothesis 1a. Although the effect of perceived control on recommendation satisfaction was found to be non-significant in the stratified model for both novices, with an estimate of 0.61 (SE = 0.32, p = 0.055) and experts, with an estimate of -0.32 (SE = 0.34, p = 0.35), it can be noted that these effects are in different directions. The result of a Sobel test, displayed in Table 5, reveals that the effects are significantly different from each

other (p = 0.046). Opposite to what was proposed in hypothesis 1b, the effect of perceived control on recommendation satisfaction is found to be more positive for novice users.

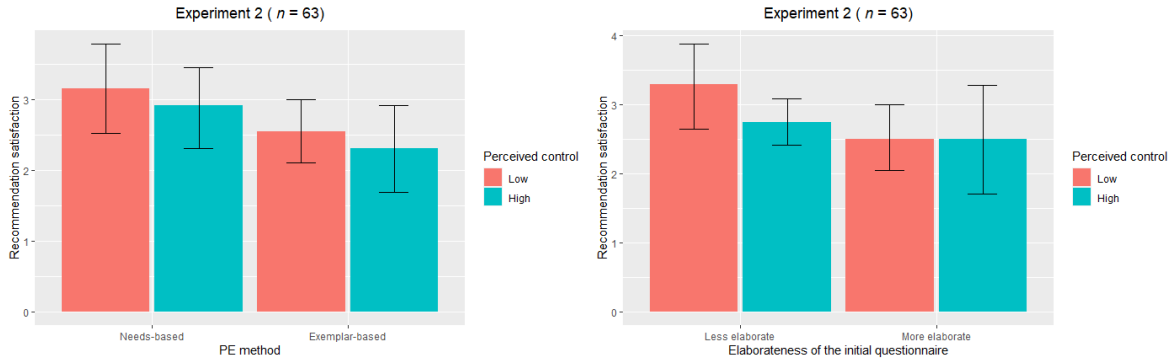

Figure 14: The mean level of recommendation satisfaction in experiment 2, displayed for each PE method (left) or level of elaborateness (right) and for each relative level of perceived control, divided by a median split.

In both plots displayed in Figure 14, there seems to be an effect of the experimental manipulations. This is also confirmed by the mediation model, as both PE method, with an estimate of -0.64 (SE = 0.31, p = 0.042), and elaborateness, with an estimate of -0.62 (SE = 0.30, p = 0.038), were found to have a negative direct effect on recommendation satisfaction. A similar negative total effect is found for both PE method, with an estimate of -0.65 (SE = 0.31, p = 0.034), and elaborateness, with an estimate of -0.60 (SE = 0.30, p = 0.049). Notably, the total and direct effects are quite similar, indicating that there is little mediation in the model. This is unsurprising, as perceived control did not affect recommendation satisfaction. In Figure 15, the mean level of perceived control is displayed for each PE method or level of elaborateness and for each level of domain knowledge. While PE method does not seem to affect perceived control, a small difference can be noted for elaborateness, with a higher level of elaborateness resulting in a higher level of perceived control. This finding is supported by the mediation model, where a positive effect of elaborateness on perceived control is found, with an estimate of 0.23 (SE = 0.12, p = 0.046). Inspecting the effects of domain knowledge using the stratified models, no significant effects are found for both novices and experts. Displayed in Table 5, a series of Sobel tests reveal that, apart from the effect of perceived control on recommendation satisfaction, the effects for novices and experts were not found to be significantly different from each other.

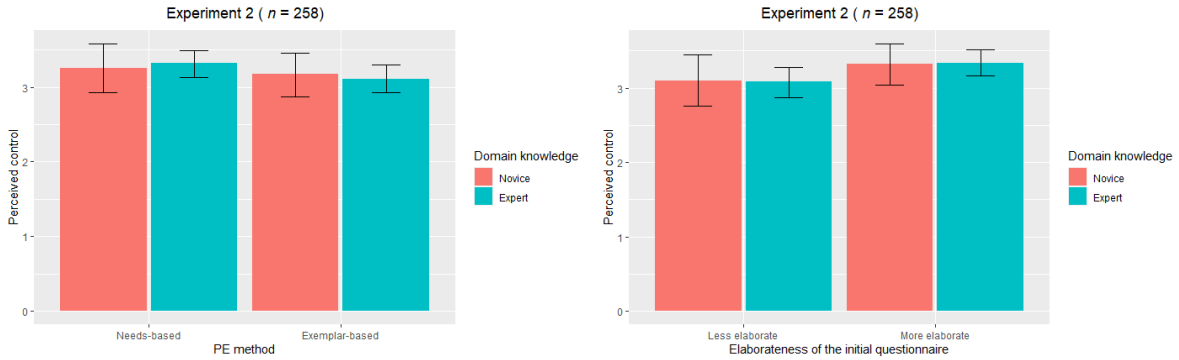

Figure 15: The mean level of perceived control in experiment 2, displayed for each PE method (left) or level of elaborateness (right) and for each level of domain knowledge.

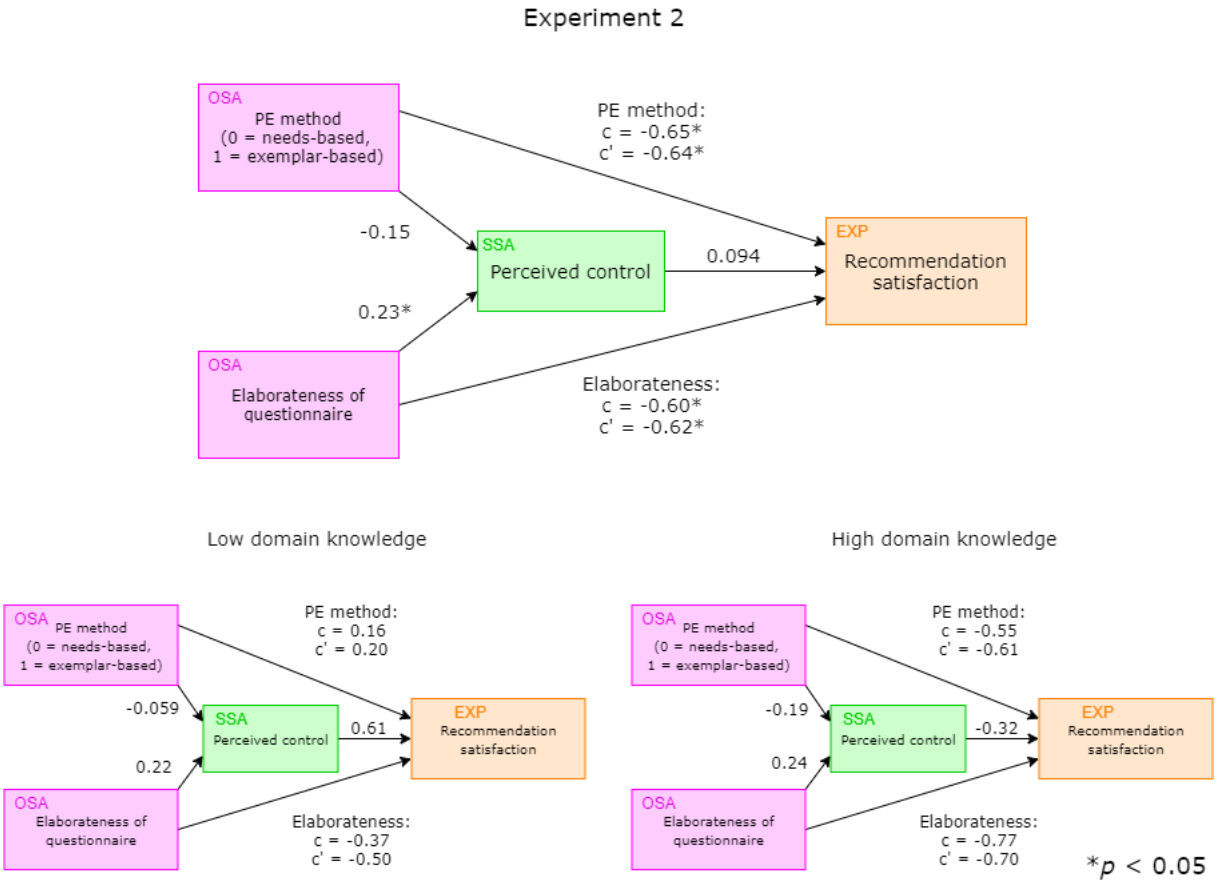

Figure 16: Results of a mediation analysis of experiment 2 (top), and the mediation analyses at each level of domain knowledge (bottom). For each effect, the regression estimate is displayed. An asterisk denotes a statistically significant effect. Appendix D displays the regression results in more detail.

In conclusion, hypothesis 2 was not supported by the second experiment, as the effects of PE method were not found to be different between novices and experts. In general, the needs-based PE method was found to produce the most satisfying recommendations. This effect was not mediated by the perceived level of control of users.

Hypothesis 3a proposed that a more elaborate questionnaire would lead to more accurate recommendations. However, the results of experiment 2 reveal an opposite effect: users who were able to change the initial weights in the recommender system were found to be less satisfied with the resulting recommendations. Hypothesis 3b was also not supported by the data, as no difference was found between the effects for novices and experts.

Effect	Estimate	*SE*	*p*
PE method			
Total effect PE method -> satisfaction	0.707	0.695	.309
Direct effect PE method -> satisfaction	0.805	0.707	.255
PE method -> perceived control	0.132	0.261	.614
Perceived control -> satisfaction	0.933	0.469	.046*
Elaborateness of the initial questionnaire			
Total effect elaborateness -> satisfaction	0.403	0.669	.547
Direct effect elaborateness -> satisfaction	0.193	0.625	.757
Elaborateness -> perceived control	-0.020	0.265	.940

Table 5: Results of Sobel tests to analyze differences of effects in experiment 2 between novices and experts. A positive estimate indicates that the effect is more positive for novices, compared to experts.

Lastly, the (stratified) mediation models of the data from experiment 2 using listwise deletion, instead of FIML, are compared with the models used in this section. Appendix E displays these results. The results are qualitatively similar. A few effects, though similar in regression estimate, are no longer significant in the model after listwise deletion, possibly due to a decrease in power resulting from the smaller sample size.

Effects on probability of acceptance

A mixed logistic regression analysis is conducted to investigate the effects of the manipulations, as well as the effect of perceived control, on the probability that a recommendation is accepted is investigated. The resulting model (n = 805, k = 258) is displayed in Figure 17. Similar to the findings of experiment 1, none of the included predictors affected the probability of acceptance, and these effects were also not found to be moderated by domain knowledge. However, the moderation of domain knowledge on the effect of PE method on the probability of acceptance, with an estimate of -0.64, was found to be

bordering on significance (*SE* = 0.33, *p* = 0.05004), indicating that the effect of an exemplar-based PE method may be more positive for novice users than for experts. This describes an opposite effect to what was proposed in hypothesis 2, which states that the effect of an exemplar-based PE method is more positive for experts.

To visualize the investigated effects, the acceptance ratio for users with different PE methods (left) or different levels of elaborateness (right), for each level of domain knowledge are displayed in Figure 18. For elaborateness, the variance in the data is deemed too large to be conclusive. For PE method, however, the acceptance ratio for experts using the exemplar-based PE method does seem to be lower than the other visualized conditions, which may explain the results of the mixed logistic regression on the probability of acceptance.

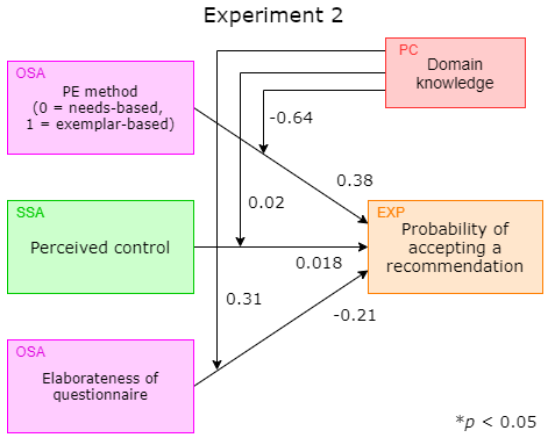

Figure 17: Logistic regression model with estimates of experiment 2. Detailed regression results are displayed in Appendix D.

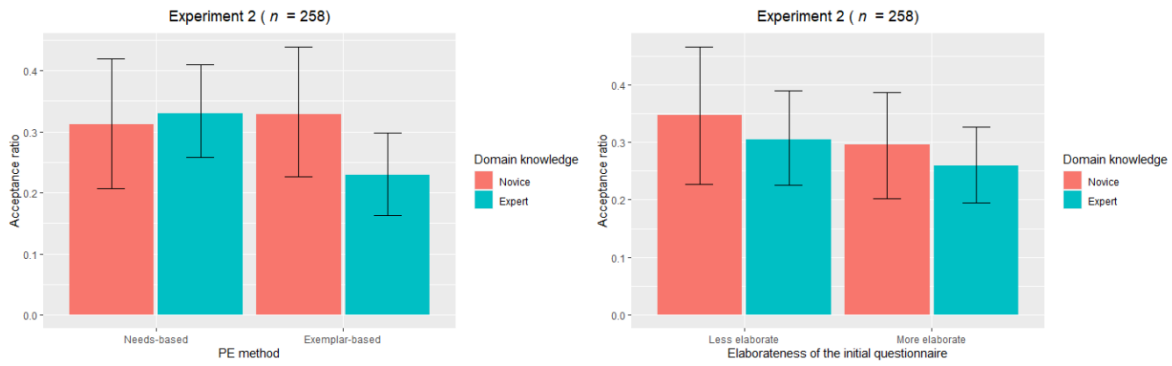

Figure 18: The mean acceptance ratio in experiment 2, displayed for each PE method (left) or level of elaborateness (right) and for each level of domain knowledge.

Lastly, a mixed logistic regression was conducted to investigate whether recommendation satisfaction also affected the probability of accepting a recommendation in experiment 2. The model (n = 276, k = 63) revealed that recommendation satisfaction was a significant predictor for the probability of acceptance, with an estimate of 0.32 (SE = 0.15, p = 0.033). Again, recommendation satisfaction, as measured in this study, cannot cause a higher probability of accepting a recommendation, as it is measured after the recommendations have been responded to. Conversely, it may be argued that having accepted more recommendations has a positive effect on the recommendation satisfaction that is indicated by the user.

Discussion

The first experiment supported the expectation that perceived control positively affects recommendation satisfaction, as posed in hypothesis 1a. As the stratified models did not show these effects for novices and experts, it cannot be concluded that this effect is stronger for experts, as proposed in hypothesis 1b. Overall, the PE method that was used did not affect users' perceived control or recommendation satisfaction. Interestingly, using either PE method did not make a difference on the produced recommendations for experts. For novices, however, the mediation analysis revealed a direct effect and an indirect effect (through perceived control) in opposite directions. These effects canceled each other out, resulting in a non-significant total effect. For novices, it is reasoned that the exemplar-based PE method is preferred over the needs-based method, except for users who are less knowledgeable regarding the attractions in the park, as this knowledge is required to operate it. Hypothesis 2 proposed that the exemplar-based PE method produced more satisfactory recommendations for experts, which was not supported by the results. Additionally, hypothesis 2 stated that the needs-based PE method produced more satisfactory recommendations for novices. It is reasoned, based on the results, that this is only found for users who lack the domain-specific knowledge required by the exemplar-based PE method. In fact, the results suggest the opposite for novice users who do possess this knowledge, for whom the exemplar-based PE method produced better recommendations.

The results of the second experiment do not support hypotheses 1a and 1b. Perceived control did not affect recommendation satisfaction and, in contrast to expectations, the effect was more positive for novice users than for experts. However, the effects for both novices and experts were not significant. Apart from this finding, none of the effects for novices and experts were different from each other, as shown by a series of Sobel tests. Hypotheses 2 and 3b, which state that the effects of an exemplar-based PE method and a more elaborate questionnaire are more positive for experts, are therefore not supported. Further, none of the investigated effects were significant in the stratified models. A reason for this may be that as the stratified models have smaller sample sizes in both experiments, they may suffer from a lack of power.

The similarity between the direct effect and the total effect for both manipulations suggests that perceived control does not mediate the main effects of the manipulation, even though a more elaborate questionnaire did positively affect perceived control. Both an exemplar-based PE method (compared to needs-based) and a more elaborate questionnaire were found to have a negative effect on users' satisfaction with the produced recommendations. This is in contrast with hypothesis 3a, which proposed

that a more elaborate questionnaire would produce more satisfying recommendations, as users were given more opportunities to personalize their profile. A possible explanation is that the two extra questions did not sufficiently personalize the recommendations, leading to disappointment for users who expected their preferences to be taken into account. Further research into users' interaction with the critiquing component is advised to explore the mechanisms underlying this effect.

In both experiments, the effects on users' probability of accepting recommendations were also measured. The formulated hypotheses were not supported by the effects on the interaction measure. Possibly, domain knowledge moderated the effect of PE method on the probability of accepting a recommendation, such that experts with an exemplar-based PE method may accept fewer recommendations. Apart from this finding, no effects were found in the logistic regression models. Although a relation is found between recommendation satisfaction and acceptance, neither the experimental manipulations nor perceived control seems to have clear effects on the interaction measure.

In conclusion, although little support is found for the hypotheses posed in this thesis, a number of interesting insights are to be obtained from the experiments presented in this thesis. Further research is needed to investigate the reproducibility of the findings and the mechanisms behind them. The effect of perceived control on recommendation satisfaction was more positive for novice users than for experts. An interpretation of these findings suggests that the effects of perceived control for people interacting with a theme park recommender system are generally quite subtle, except when a user does not possess the knowledge required to answer the questions asked. In that case, recommendation satisfaction is affected, as the user has less opportunity to disclose their personal preferences. This is supported by the additional findings that the effects of perceived control on recommendation satisfaction were, in both experiments, closer to 0 for experts. It can be expected that expert users usually do possess the knowledge required to answer the questions in the questionnaire. When controlling for the level of perceived control, however, novice users were generally more satisfied with the recommendations that were produced by the exemplar-based PE method. A suitable solution for theme park recommender systems may be to implement a PE method that provides control over domain-specific aspects of the recommendations, while also giving the option to indicate a lack of knowledge regarding the questions. This way, the system can allow users who are unfamiliar with the domain to request more indirect ways of indicating preferences, such as the needs-based PE method described in this thesis.

Lastly, while adding a critiquing component did improve users' interaction with the recommender system, as reflected by a higher acceptance ratio in the second experiment and a possible higher level of recommendation satisfaction, the more elaborate questionnaire did not. Users with a more elaborate questionnaire did perceive a higher level of control, but they were less satisfied with the resulting recommendations. Although further research is recommended into the mechanics behind the effects, a possible reason could be that the more elaborate questionnaire did not result in the personalization that people expected from it, which may have led to disappointment.

Limitations and suggestions for future research

The study presented in this thesis is subject to several limitations, discussed in this section, that need to be taken into consideration when interpreting its results. Future research could attempt to reproduce the findings presented, while mitigating these limitations, to investigate the validity and reproducibility of the research presented in this thesis.

COVID-19

The most prominent limitations to the study presented in this thesis are caused by the COVID-19 pandemic, during which the study was conducted. The emergence of the pandemic had several unforeseen implications for Efteling. The closing of the park, requiring guests to register their visit online, and informing visitors of additional safety measures to allow for social distancing, were a few of the aspects that were prioritized by the development team over the development of the recommender system. The recommender system was originally planned to be released around June of 2020. The implications of the pandemic, in combination with other underestimations of the time required for the development, delayed this planning. The public release date of the application was four months later and after the summer holidays, on October 20th. Due to this severe delay, relatively little data could be collected for this study. Additionally, the effects of COVID-19 forced Efteling to close again from the 15th of December, eliminating the opportunity to collect data during the holiday season. The small temporal window also removed the opportunity to conduct a study with an iterative approach, applying insights from early experiments.

Aside from the quantity of the collected data, the COVID-19 pandemic may also have repercussions for the validity and generalizability of the data that was gathered. The situation was unprecedented for Efteling, so it is difficult to estimate how it affected guests who visited the park during the pandemic. It can be expected that the emphasis on social distancing influences, for example, visitors' perceptions of attraction lines and make them more averse to crowded areas. As the recommendation engine had been trained using data that was gathered before the pandemic, the pandemic might have impacted

the accuracy of the recommendations. Understanding the generalizability of the results, therefore, requires analysis of the high-level changes in the attitude of visitors during this time, compared to attitudes before the start of the pandemic, which is outside the scope of the presented study.

As a more practical limitation, the experiments were conducted on the recommender system when it only recently had been deployed. The first experiment started one month after public release. During this month, the park was closed for two weeks due to government measures to mitigate the spread of the COVID-19 virus. Therefore, a well-documented baseline of users' experience with the system and its prediction accuracy had not been established yet. Not only had certain features not been included yet that could accelerate data collection, such as a version for Android devices and multilingual services, but a few bugs may also still have persisted in the system. It is impossible to identify data points of users who experienced technical difficulties, yet they might have impacted the satisfaction ratings of users. It is expected that, on a large scale, this influence would be negligible, but anomalies in the data could affect the results of analyses on a dataset of a smaller scale. However, due to the random distribution of participants over the conditions, this is not expected to heavily affect the manipulation effects. In addition, the constructs that were analyzed in this study were measured using single questionnaire items. As the study was conducted on a live product, including large item-batteries for each construct was infeasible with respect to the usability of Pennenveer. Due to this, it is not certain whether the questionnaire items reflect the measured constructs in an optimal way. Again, when a large number of data points would be collected, the sample size could compensate for measures that were less than perfect. With a smaller dataset, it is advised to further research the validity of the measures in a controlled environment. A smaller user trial could be conducted to investigate the optimal way to measure the constructs under investigation.

In general, it is encouraged to conduct more elaborate research on a larger sample, in a less volatile societal climate, to verify the generalizability of the findings and to investigate whether other effects or underlying mechanisms can be identified.

Other limitations

A number of other limitations can be identified that could have affected the results of the study. First of all, there was a relatively small number of users who rarely visited Efteling. About a third of the users in both datasets were indicated as novices, and two-thirds were experts. Note that the distinction between experts and novices was made such that it provided the most even split. It is not surprising that frequent Efteling visitors are more willing to download the application out of curiosity for the newest service. To conduct a thorough analysis regarding the effects of domain knowledge, future research should attempt to include more inexperienced visitors, by adding, for example, multilingual

services to the recommender system. It can also be expected that the COVID-19 pandemic influenced the amount of long-distance tourism that Efteling generated, increasing the share of visitors who own a season pass or live close to the park.

Another limitation in the data is the lack of observations gathered for recommendation satisfaction. This construct was measured as part of the final questionnaire, which most people neglected to fill in. Apart from severely limiting the proportion of complete data, a bias may be present in the data. The analysis methods described in this thesis are based on the assumption that missing data is nonignorable, meaning that the reason for missingness does not depend on the value of the observation. It is impossible to determine whether missing data is nonignorable without conducting follow-up research on the users who produced the data. It is not unthinkable, however, that there is a share of users who did not fill in the final questionnaire, simply because they were unsatisfied with the recommendations and stopped using the application. If this were the case, a share of unsatisfied users would not be included in the data. This would mean that the sample of users that did fill in the final questionnaire would systematically have higher values for recommendation satisfaction than the true population. In this thesis, Full Information Maximum Likelihood was used to retain as much information in the models as possible. If missing data would be nonignorable, this would imply that the effects of the manipulations on perceived control would not describe the same population as the effects on recommendation satisfaction, weakening the validity of the path analyses. The alternative considered method of dealing with missingness is listwise deletion, the results of which are displayed in Appendix E. If missing data would be nonignorable, listwise deletion would produce results describing effects for users with a relatively high level of recommendation satisfaction, instead of the true population. Further research should attempt to persuade more users to fill in the final questionnaire to diminish a potential bias and compare different methods of measuring recommendation satisfaction, to avoid nonignorable missing data altogether.

Lastly, during the two-week closure of the park in the month prior to the first experiment, eight instances were logged of users who received a recommendation. This data was not included in the dataset but does exemplify that illegitimate data may also have been generated outside this time period. In general, recommendations can only be requested by users when they are inside the confinements of Efteling. However, it is possible to fake a GPS location, thereby prompting a recommendation while not in the park. This data would not be distinguishable from legitimate data in the collected dataset. Again, on a large-scale dataset, these interferences are estimated to be negligible, but they may affect the results of a smaller dataset.

Future research directions

In addition to mitigating the limitations described earlier, such as the quantity of the collected data and the share of users with a low level of domain knowledge, there are some other directions for future research that are identified based on this thesis. An important factor to consider is that the theme park recommender system is rather unique in its implementation, so many insights can be gathered from investigating various aspects of the platform, and manipulations on it.

Firstly, further research could explore more ways of implementing preference elicitation methods in the recommender system. In the presented thesis, a needs-based and an exemplar-based PE method have been compared. Further research could investigate how different PE methods affect user experience, but also how these PE methods can be improved or even combined to accommodate various theme park visitors. An example provided in this thesis is giving users an exemplar-based PE method, with the option to indicate that they do not know the attractions, upon which they are given a needs-based PE method. Further, a qualitative study could be conducted to identify the needs and expectations of various users, to find ways to accommodate these.

Also, in this thesis, visiting frequency of Efteling was measured as a proxy for domain knowledge. However, further research could distinguish between the amount of experience users have with Efteling specifically, and with theme parks in general. For example, a theme park enthusiast who has never visited Efteling before may not feel comfortable indicating preferences regarding specific attractions but may still want to fine-tune other settings to make the most of their visit. A user who has rarely visited any theme park, on the other hand, could not know what to expect and prefer to be shown around.

Additionally, more ways of implementing the critiquing component can be compared. The changing of the weights as it has been implemented now has yet to be validated. Further research could collect data regarding the critiquing behavior of users, to investigate, for example, which critiques users want to give, how often they want to use it, and whether it improves the probability of accepting the recommendations that follow. This way, the optimal critiquing options to include and the optimal values for changing the weights in the engine can be validated.

Recommendations for Efteling

Aside from general suggestions for future research in the domain, a number of suggestions are provided for the further development of Pennenveer that may be of interest. Firstly, although the recommender system was found to perform better in the second experiment than in the first, recommendation satisfaction is considered quite low (M = 2.76, SE = 0.16), given Efteling's aspirations of providing a 9+ experience (out of 10) for all visitors. This may indicate that there are still a number of pressing issues

that diminish users' experience with Pennenveer. It is suggested to perform a qualitative study, by inspecting the feedback that was provided in the application or by interviewing application users with varying levels of domain knowledge, to uncover ways of improving the general satisfaction of users of Pennenveer.

Secondly, the results of experiment 2 suggest that the needs-based PE method produced more satisfactory recommendations than the exemplar-based PE method. However, the first experiment showed that the exemplar-based PE method was preferred for novices. The opposite was found when those users perceived a low level of control, which may be caused by an experienced difficulty to indicate preferences for specific attractions for users that are not familiar with the attractions. For experts in experiment 1, this effect was not found. To further investigate the mechanisms behind the effects of varying PE methods, a qualitative study could be conducted to assess how people interact with the PE methods. To accommodate users that are not familiar with the attractions, while still taking advantage of a higher level of satisfaction produced by the exemplar-based PE method for novices, further research could also investigate a combination of the two PE methods described in this thesis. A way to achieve this would be to display the exemplar-based PE method, along with a way for users to indicate that they are unfamiliar with the attractions and allowing them to switch to a needs-based PE method. This may increase perceived control, without forcing users to answer questions that they may not know the answer to.

Lastly, in experiment 2, the elaborateness of the initial questionnaire was found to negatively influence recommendation satisfaction. The more elaborate questionnaire allowed users to manipulate the initial weight of certain aspects of the recommendations and was expected to positively influence satisfaction. A possible reason for this finding is that the additional questions in the more elaborate questionnaire raised the expectation of users regarding the personalization of the recommendations. If these expectations were then not met, this could lead to disappointment. However, the way that critiquing data was collected for this study did not make it possible to inspect details of the critiques, such as which critique is given for each recommendation and how the values of the various weights affected the recommendations that were provided. Therefore, it is only possible to measure the outcome effects of implementing the critiquing component, without validating the underlying mechanisms that cause them. Storing these interactions would make it possible to further analyze the optimal method and magnitude of altering the weights in the recommendation engine. For example, further research could explore the extent to which a higher value for a certain weight actually resulted in the associated feature having a larger influence on the recommendations.

Conclusion

This thesis describes a study that has been conducted alongside the deployment of a conversational recommender system at the Dutch theme park Efteling. Several factors are expected to impact the generalizability and validity of the results, many of which are caused by the COVID-19 pandemic. Due to this unfortunate timing, it is strongly encouraged to continue researching these effects and attempt to unveil underlying mechanisms that might explain the effects found. Additional directions for future research are discussed. Although the hypotheses posed in the presented study are largely unsupported by the data, there are a few insights that are worth considering.

A user's satisfaction with the recommendations that are provided by a theme park recommender system may be affected by their perceived level of control. This effect is mainly present for users with a low level of domain knowledge. This might be because less experienced users have less opportunity to indicate their preferences when they lack the domain-specific knowledge to answer the questions asked. The reasoning is supported by findings regarding the PE method that is used. Neither of the PE methods, which differed in the amount of domain-specific knowledge they required, generally resulted in more satisfactory recommendations than the other. However, for relatively inexperienced users, it seems that the PE method that required more domain-specific knowledge was preferred, except for users who lacked the knowledge to answer the question. For experienced users, both PE methods performed similarly, suggesting that their satisfaction with the recommendations was not affected by the amount of domain-specific control they were offered. Lastly, a more elaborate questionnaire, which allowed users to indicate additional preferences to further personalize their recommendations, negatively affected users' satisfaction with these recommendations. The extra questions did, however, positively affect perceived control. Although further research is required, the additional questions might not have personalized the recommendations as much as users expected, which could have led to disappointment.

Despite a fair number of limitations on the expected validity of the presented study, support is found for the notion that perceived controllability affects users' satisfaction with the recommendations. Additionally, the optimal level of domain-specific control that a PE method offers the user, does seem to depend on the amount of knowledge that the user has about the domain, in line with literature. Further uncovering differences between users with varying levels of domain knowledge is considered worthwhile. Exploring the nature of these differences, for example by conducting qualitative research or investigating the mechanisms behind certain effects, is an important step towards ensuring a magical experience for every user.

Literature

Abbaspourghomi, A. (2020). *A personalized recommendation system for Efteling using crowdedness and guest behavior* (Unpublished PDEng report). Eindhoven University of Technology, the Netherlands.

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, *17*(6), 734-749. https://doi.org/10.1109/TKDE.2005.99

Balabanović, M., & Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM*, *40*(3), 66-72. https://doi.org/10.1145/245108.245124

Bettman, J. R., Luce, M. F., & Payne, J. W. (1998). Constructive consumer choice processes. *Journal of consumer research*, *25*(3), 187-217. https://doi.org/10.1086/209535

Bobadilla, J., Ortega, F., Hernando, A., & Bernal, J. (2012). A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-based systems*, *26*, 225-238. https://doi.org/10.1016/j.knosys.2011.07.021

Bostandjiev, S., O'Donovan, J., & Höllerer, T. (2012). TasteWeights: a visual interactive hybrid recommender system. In *Proceedings of the sixth ACM conference on Recommender systems*, 35-42. https://doi.org/10.1145/2365952.2365964

Burke, R. (2007). Hybrid web recommender systems. In *The adaptive web* (pp. 377-408). Springer, Berlin, Heidelberg.

Christakopoulou, K., Radlinski, F., & Hofmann, K. (2016, August). Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,* 815-824. https://doi.org/10.1145/2939672.2939746

Cherrie, M. (2019, January 17). *How to build structural equation model in Lavaan*. Retrieved from https://www.markcherrie.net/post/structural-equation-modelling-in-r/

Efteling. (2019). *De visie van de Efteling op 2030.* Retrieved from Efteling website: https://www.efteling.com/nl/pers/visie-2030/

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological methods*, *6*(4), 352-370. https://doi.org/10.1037/1082-989X.6.4.352

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural equation modeling*, *8*(3), 430-457.

Felfernig, A., & Burke, R. (2008, August). Constraint-based recommender systems: technologies and research issues. In *Proceedings of the 10th international conference on Electronic commerce* (pp. 1-10). https://doi.org/10.1145/1409540.1409544

Häubl, G., & Trifts, V. (2000). Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing science*, *19*(1), 4-21. https://doi.org/10.1287/mksc.19.1.4.15178

Hu, R., & Pu, P. (2009). A comparative user study on rating vs. personality quiz based preference elicitation methods. In *Proceedings of the 14th international conference on Intelligent user interfaces*, 367-372. https://doi.org/10.1145/1502650.1502702

Hutton, R. J., & Klein, G. (1999). Expert decision making. *Systems Engineering: The Journal of The International Council on Systems Engineering*, *2*(1), 32-45. https://doi.org/10.1002/(SICI)1520-6858(1999)2:1<32::AID-SYS3>3.0.CO;2-P

He, C., Parra, D., & Verbert, K. (2016). Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications*, *56*, 9-27. https://doi.org/10.1016/j.eswa.2016.02.013

Gretzel, U. (2011). Intelligent systems in tourism: A social science perspective. *Annals of tourism research*, *38*(3), 757-779. https://doi.org/10.1016/j.annals.2011.04.014

Jannach, D., Manzoor, A., Cai, W., & Chen, L. (2020). *A Survey on Conversational Recommender Systems.* Manuscript submitted for publication. Retrieved from arXiv:2004.00646

Jugovac, M., & Jannach, D. (2017). Interacting with recommenders—overview and research directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *7*(3), 1-46. https://doi.org/10.1145/3001837

Knijnenburg, B.P., Reijmer, N.J., & Willemsen, M.C. (2011). Each to his own: How different users call for different interaction methods in recommender systems. In *RecSys'11 - Proceedings of the 5th ACM Conference on Recommender Systems*, 141-148. https://doi.org/10.1145/2043932.2043960

Knijnenburg, B. P., & Willemsen, M. C. (2015). Evaluating recommender systems with user experiments. In *Recommender Systems Handbook,* 309-352. Springer, Boston, MA.

Knijnenburg, B.P., Willemsen, M.C., & Broeders, R. (2014). Smart sustainability through System Satisfaction: Tailored Preference Elicitation for Energy-saving Recommenders. *Full paper accepted to the Americas Conference on Information Systems (AMCIS).*

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, *22*(4-5), 441-504. https://doi.org/10.1007/s11257-011-9118-4

Knijnenburg, B.P., Willemsen, M.C., & Kobsa, A. (2011). A pragmatic procedure to support the user-centric evaluation of recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems,* 321-324. https://doi.org/10.1145/2043932.2043993

Lathia, N. (2015). The anatomy of mobile location-based recommender systems. In *Recommender Systems Handbook* (pp. 493-510). Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7637-6_14

Li, W. L. (2010, December). Impact of waiting time on tourists satisfaction in a theme park: An empirical investigation. In *2010 IEEE International Conference on Industrial Engineering and Engineering Management* (pp. 434-437). IEEE. https://doi.org/10.1109/IEEM.2010.5674481

Lu, J., Wu, D., Mao, M., Wang, W., & Zhang, G. (2015). Recommender system application developments: a survey. *Decision Support Systems*, *74*, 12-32. https://doi.org/10.1016/j.dss.2015.03.008

Nguyen, T.N., & Ricci, F. (2018). A chat-based group recommender system for tourism. *Information Technology & Tourism*, *18*(1-4), 5-28. https://doi.org/10.1007/s40558-017-0099-y

Orense, B., Chandrasekaran, C., Yolande, E., & Rashipour, Z. (2020). *Developing a monitoring and information tool for identification of guest profiles* (Unpublished PDEng report). Eindhoven University of Technology, the Netherlands.

Payne, J.W., Bettman, J.R., Schkade, D.A., Schwarz, N., & Gregory, R. (1999). Measuring constructed preferences: Towards a building code. In Fischhoff, B., Manski, C.F. (Eds.), *Elicitation of preferences* (pp. 243-275). Springer, Dordrecht. https://doi.org/10.1007/978-94-017-1406-8_9

Pommeranz, A., Broekens, J., Wiggers, P., Brinkman, W.P., & Jonker, C.M. (2012). Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process. *User Modeling and User-Adapted Interaction*, *22*(4-5), 357-397. https://doi.org/10.1007/s11257-011-9116-6

Pu, P., Chen, L., & Hu, R. (2012). Evaluating recommender systems from the user's perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction*, *22*(4-5), 317-355. https://doi.org/10.1007/s11257-011-9115-7

Randall, T., Terwiesch, C., & Ulrich, K.T. (2007). Research note—user design of customized products. *Marketing Science*, *26*(2), 268-280. https://doi.org/10.1287/mksc.1050.0116

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, *40*(3), 56-58. https://doi.org/10.1145/245108.245121

Themed Entertainment Association & AECOM. (2019). *TEA/AECOM 2019 Theme Index and Museum Index: The Global Attractions Attendance Report.* Retrieved from AECOM website: https://aecom.com/content/wp-content/uploads/2020/07/Theme-Index-2019.pdf

Appendix A

Questions included in the initial and final questionnaire of Pennenveer

Initial questionnaire

#	Question type	Question	Answer Options
1	The user can click on any age group and input the respective amount of people in their company.	Wat is de samenstelling van je gezelschap? Selecteer het aantal personen per leeftijdscategorie..	0-4 Years old 5-9 Years old 10-17 Years old 18-29 Years old 30-49 Years old 50+ Years old
2	Multiple choice	Hoe vaak bezoek je gemiddeld de Efteling?	Eens per halfjaar of vaker Eens per jaar Eens per 2 jaar Eens per 3 jaar of minder vaak Dit is mijn eerste bezoek aan de Efteling
3a	User can select up to five attractions. [This Question is part of experiment 1 and 2: 50% of users receive this question, the other half will be asked question 3b]	Dit is mijn favoriete vraag: welke attracties willen jullie vandaag het allerliefste bezoeken? Je kunt er tussen de 1 en de 5 kiezen  .	Baron 1898 Symbolica Max & Moritz Fabula Python Carnaval Festival Vogel Rok Sprookjesbos Joris en de Draak
	..	..	..

	..	..	..
3b	This question consists of two multiple choice questions. [This Question is part of experiment 1 and 2: 50% of users receive this question, the other half will be asked question 3a]	Deze is leuk! Met deze vragen kan ik bepalen wat voor attracties jullie niet mogen missen.	Hoe dapper is jullie gezelschap? '😟 Wij doen liever niet al te spannende dingen...' '😁 Wij doen het rustig aan, maar soms willen we in een spannende attractie...' '😎 Wij zijn echte helden! We zijn nergens bang voor...' Hoe graag willen jullie wegdromen bij alle verhalen? '🎠 Wij komen vooral voor alle leuke attracties!' '🏰 Wij kiezen voor een combinatie van gave attracties en prachtige verhalen!' '✨ Wij laten ons meeslepen in alle betoverende verhalen!'
4	This question consists of two multiple choice questions. [This Question is part of experiment 2: 50% of users receive this question, the other half will not]	Hiermee kan ik nog beter inschatten wanneer het ideale moment is om een attractie aan te bevelen!	Hoe ver willen jullie lopen voor een leuke attractie? '🚂 Ik vind het niet erg om een eindje te lopen!' '🌳 Ik wil best een stukje lopen, maar niet te ver..' '📌 Ik wil het liefst in attracties die vlakbij zijn.' Hoe erg vinden jullie het om in een wachtrij te staan? '🦥 Niet zo erg, lekker spanning opbouwen!' '⏱️ Ik heb wel geduld, maar niet te lang!' '🎟️ Ik kies voor attracties waar ik zo snel mogelijk in kan!'
5	Multiple choice	Nog één vraag: Hoeveel invloed had je voor je gevoel over het aangeven van welke attracties je graag wil bezoeken?	Heel weinig Weinig Niet weinig en niet veel Veel Heel veel

Table A1: Contents of the questions asked in the initial questionnaire included in Pennenveer.

Final questionnaire

#	Question type	Question	Answer Options
1	Star rating (1-5)	Hoe tevreden ben je met de adviezen voor attracties die Pennenveer je heeft gegeven?	Not applicable
2	Star rating (1-5)	Hoe goed heeft de restaurant-aanbeveling je geholpen om het juiste restaurant te vinden?	Not applicable
3	Star rating (1-5)	Hoe beoordeel je restaurant {}?	Not applicable
4	Open question	Welke horeca locatie heb je vandaag wel bezocht?	Not applicable
5	Star rating (1-5)	Hoe beoordeel je die locatie?	Not applicable
6	Multiple choice	Wat heeft je gezelschap in totaal ongeveer uitgegeven aan eten en drinken vandaag?	€0,- tot €9,99 €10,- tot €19,99 €20,- tot €39,99,- €40,- tot €69,99,- €70,- tot €100,- Meer dan €100,- Dat zeg ik liever niet
7	Multiple choice	Hoe druk vond je het op jouw bezoekdag?	Heel rustig Rustig Niet rustig en niet druk Druk Heel druk
8	Star rating (1-5)	Hoe leuk vond jij jouw Efteling-bezoek?	Not applicable
9	Multiple choice	Heb je een Efteling abonnement?	Ja Nee
10	Multiple choice	Verblijf je in één van onze mooie hotels of vakantieparken?	Ja Nee
11	Open question	Hoe kunnen we onze adviezen nog beter of leuker maken?	Not applicable

Table A2: Contents of the questions asked in the final questionnaire included in Pennenveer.

Appendix B

Screenshots of the user interface of Pennenveer

This Appendix displays various sections of Pennenveer, to visualize the representation of the questions and the recommendations in the application. Note that the aim of this appendix is to provide a representative overview of the user interface of Pennenveer, and therefore does not include every question in the questionnaires. A comprehensive list of questionnaire items included in Pennenveer is provided in Appendix A.

Initial questionnaire

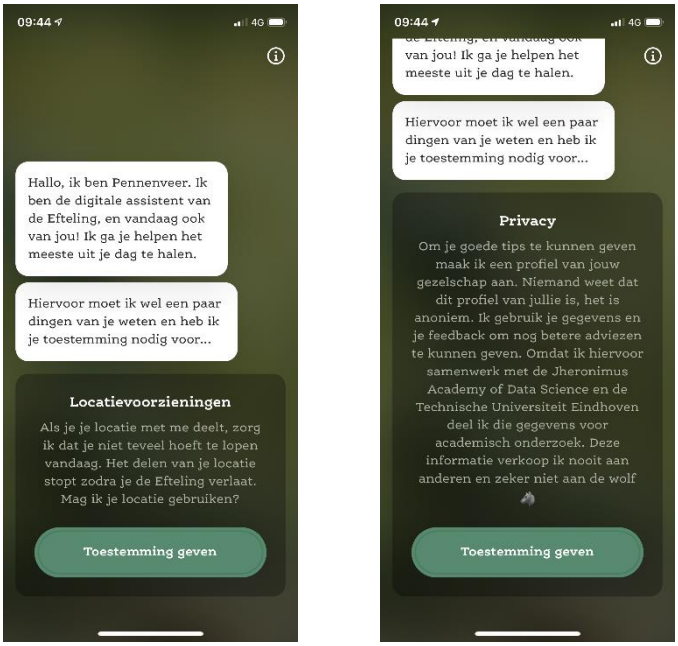

Figure B1: Initial permissions upon starting Pennenveer after installation. In addition, permission for notifications are requested.

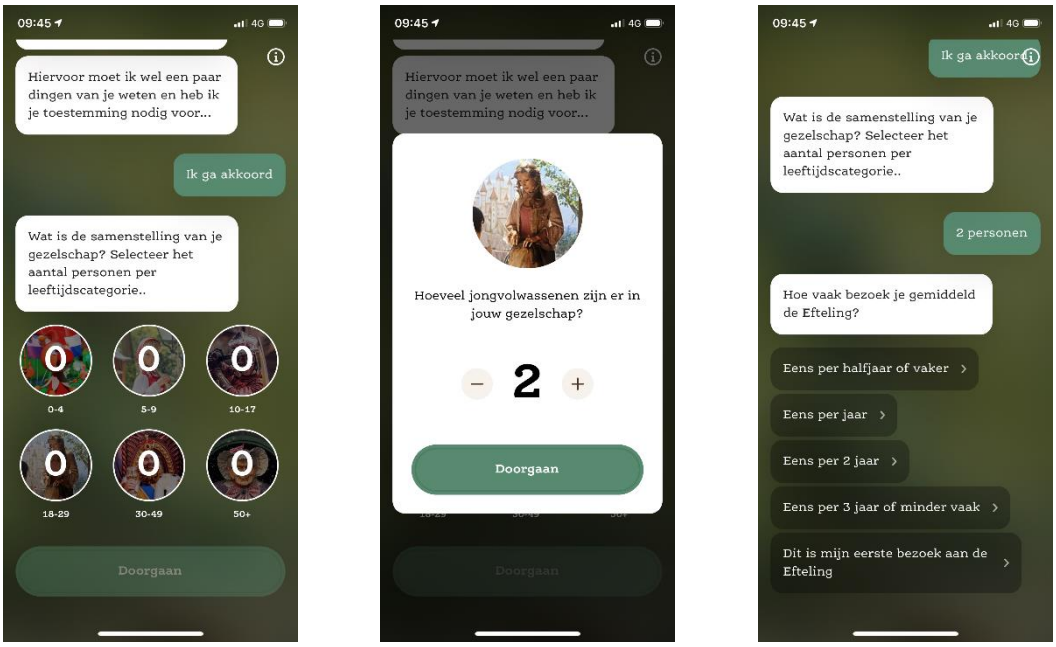

Figure B2: After granting permissions, the group composition is asked (left). Users can tap each age bucket and indicate the amount of group members within that age bucket (center). After this, the visiting frequency of the user is asked (right).

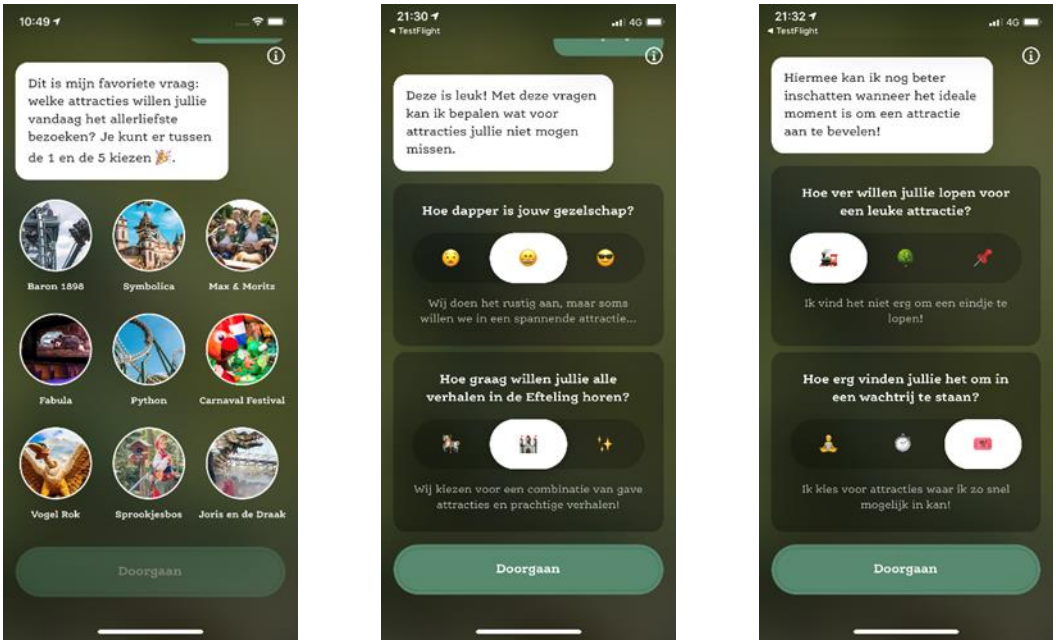

Figure B3: Experimental manipulations. In both experiments, using A/B testing, half of the users received an exemplar-based PE method (left) and half of the users received a needs-based PE method (center). After this, half of the users in experiment 2 received questions regarding their preferences of waiting times and walking distance (right).

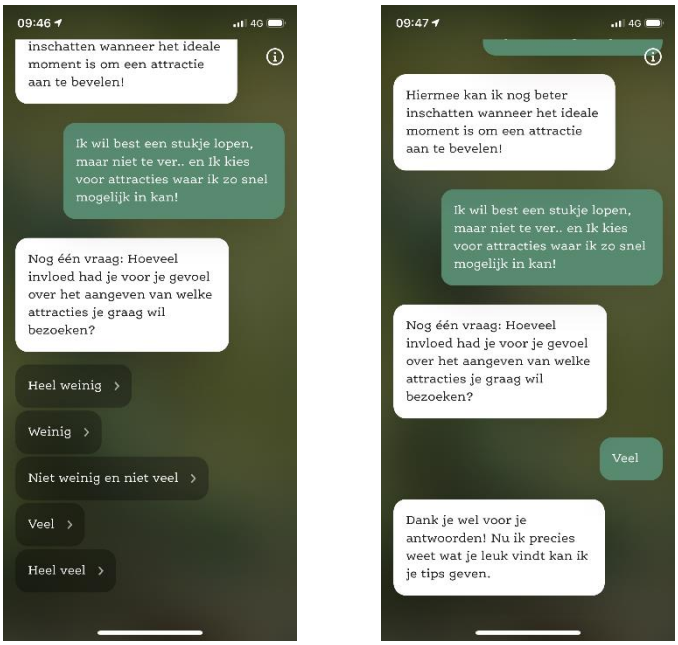

Figure B4: After the experimental manipulations, the user's perceived control is measured (left), after which the initial questionnaire is completed (right).

Recommendations

Figure B5: A recommendation is shown as a suggestion for a single attraction, along with an image, the attraction type, and the current waiting time at the attraction.

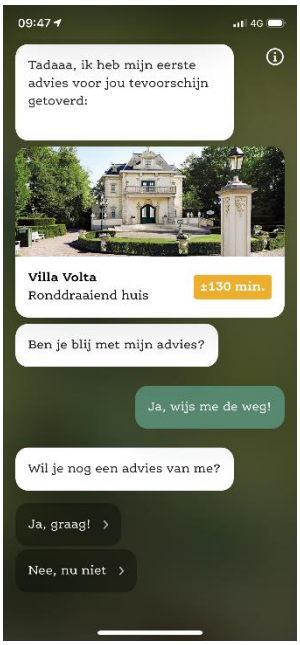

Figure B6: Upon accepting a recommendation, Pennenveer automatically opens the main Efteling-app on the page of the recommended attraction to allow navigating there immediately. After this, a new recommendation can be requested.

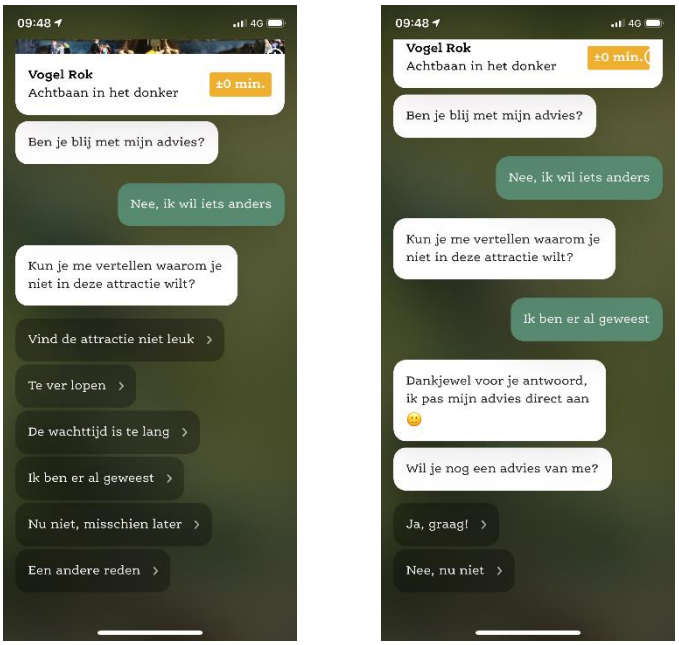

Figure B7: Upon rejecting a recommendation, the user is asked to provide feedback for their reason of rejection (left). After answering the question, the user can choose to request a new recommendation immediately or pause the conversation for the time being (right). Note that in experiment 1, the user was informed that their feedback was not processed immediately.

Final questionnaire

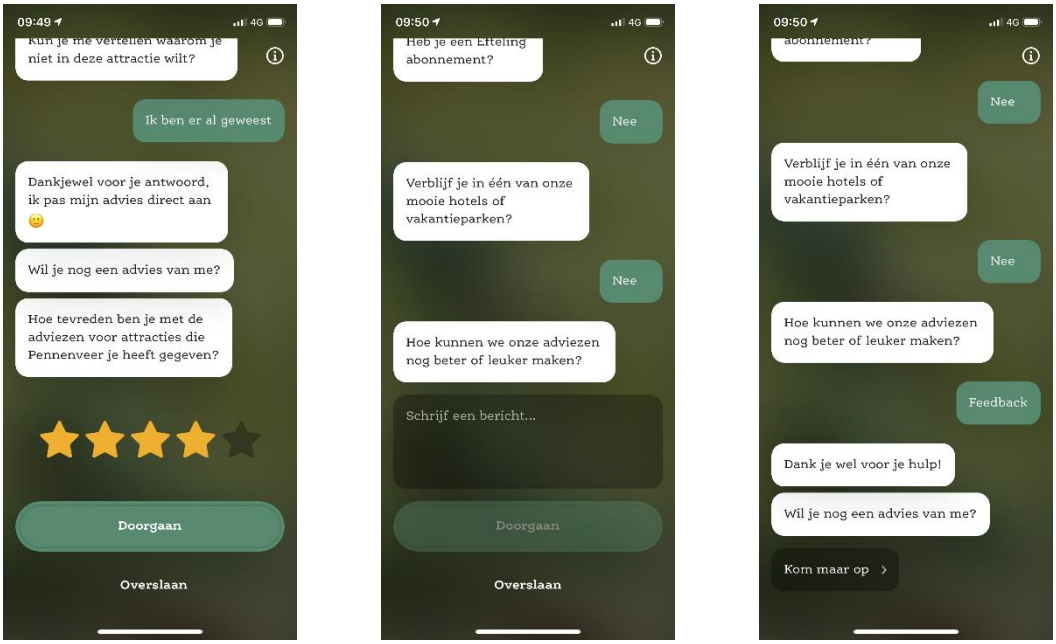

Figure B8: The first question of the final questionnaire measures recommendation satisfaction using a star rating (left). After several other questions, the last question allows users to freely provide additional textual feedback (center). When the last question is answered, the user is thanked for their effort. After this, they are still able to receive new recommendations, provided that they are still in the park (right).

Appendix C

Conversion of importance indications to attraction preferences in the needs-based PE method

		Need for Storytelling		
		Low	Medium	High
Need for Thrill	Low	Stoomcarrousel	Fairytale Forest	Fairytale Forest
		De Oude Tufferbaan	Symbolica	Symbolica
		Monorail	Fata Morgana	Fata Morgana
		Kinderspoor	Droomvlucht	Droomvlucht
		Polka Marina	Carnaval Festival	Fabula
	Medium	Pagode	Villa Volta	Fairytale Forest
		Vogel Rok	Joris en de Draak	Symbolica
		Piraña	Vliegende Hollander	Vliegende Hollander
		Monsieur Cannibale	Max & Moritz	Villa Volta
		Joris en de Draak	Symbolica	Droomvlucht
	High	Baron 1898	Baron 1898	Baron 1898
		Python	Python	Vliegende Hollander
		Joris en de Draak	Joris en de Draak	Villa Volta
		Vogel Rok	Villa Volta	Fairytale Forest
		Halve Maen	Vliegende Hollander	Droomvlucht

Table C1: Conversion table from the user's indicated needs on two dimensions to five specific attraction preferences.

Appendix D

Regression tables

Experiment 1

	Full sample			Low domain knowledge			High domain knowledge		
Effect	*Estimate*	*SE*	*p*	*Estimate*	*SE*	*p*	*Estimate*	*SE*	*p*
Total effect PE method -> satisfaction	0.066	0.330	.841	0.647	0.416	.120	-0.189	0.444	.671
Direct effect PE method -> satisfaction	0.133	0.333	.689	0.958	0.461	.037*	-0.194	0.442	.661
PE method -> perceived control	-0.171	0.129	.186	-0.533	0.188	.004*	0.024	0.159	.880
Perceived control -> satisfaction	0.394	0.190	.038*	0.583	0.353	.099	0.206	0.267	.440
Indirect effect PE method -> satisfaction	-0.067	0.062	.280	-0.311	0.222	.161	0.005	0.056	.929

Table D1: Regression table of the mediation model for the effects on recommendation satisfaction in experiment 1, including stratified models with both levels of domain knowledge.

Predictor	***Estimate***	***SE***	***p***
PE method	0.110	0.433	.799
PE method * domain knowledge	-0.391	0.513	.445
Perceived control	-0.133	0.171	.438
Perceived control * domain knowledge	0.081	0.116	.482

Table D2: Regression table of the mixed logistic regression model for the effects on probability of acceptance in experiment 1, including interaction effects.

Experiment 2

	Full sample			Low domain knowledge			High domain knowledge		
Effect	*Estimate*	*SE*	*p*	*Estimate*	*SE*	*p*	*Estimate*	*SE*	*p*
Total effect PE method -> satisfaction	-0.651	0.307	.034*	0.162	0.593	.785	-0.545	0.362	.132
Direct effect PE method -> satisfaction	-0.638	0.313	.042*	0.198	0.623	.750	-0.606	0.345	.079
Total effect elaborateness -> satisfaction	-0.600	0.304	.049*	-0.369	0.528	.484	-0.772	0.407	.058
Direct effect elaborateness -> satisfaction	-0.622	0.299	.038*	-0.503	0.503	.318	-0.696	0.365	.056
PE method -> perceived control	-0.146	0.121	.227	-0.059	0.222	.789	-0.191	0.141	.174
Elaborateness -> perceived control	0.230	0.116	.046*	0.218	0.218	.319	0.238	0.142	.094
Perceived control -> satisfaction	0.094	0.188	.616	0.614	0.320	.055	-0.320	0.342	.350
Indirect effect PE method -> perceived control	-0.014	0.038	.719	-0.036	0.174	.834	0.061	0.099	.539
Indirect effect elaborateness -> perceived control	0.022	0.049	.656	0.134	0.170	.431	-0.076	0.097	.431

Table D3: Regression table of the mediation model for the effects on recommendation satisfaction in experiment 2, including stratified models with both levels of domain knowledge.

Predictor	***Estimate***	***SE***	***p***
PE method	0.377	0.257	.142
PE method * domain knowledge	-0.637	0.325	.050
Elaborateness	-0.206	0.267	.439
Elaborateness * domain knowledge	0.314	0.347	.365
Perceived control	0.018	0.109	.867
Perceived control * domain knowledge	0.024	0.087	.781

Table D4: Regression table of the mixed logistic regression model for the effects on probability of acceptance in experiment 2, including interaction effects.

Appendix E

Path analysis results with listwise deletion as missing data method

Experiment 1

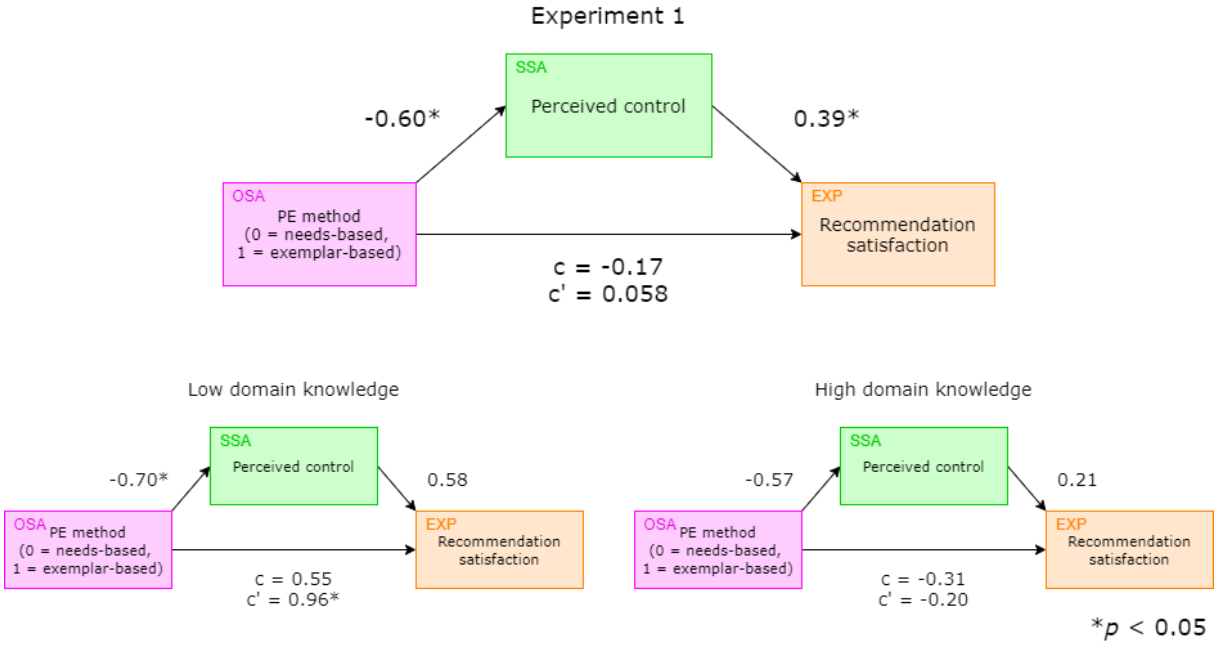

Figure E1: Results of a mediation analysis of experiment 1 (top), and the mediation analyses at each level of domain knowledge (bottom). For each effect, the regression estimate is displayed. An asterisk denotes a statistically significant effect.

	Full sample			Low domain knowledge			High domain knowledge		
Effect	*Estimate*	*SE*	*p*	*Estimate*	*SE*	*p*	*Estimate*	*SE*	*p*
Total effect PE method -> satisfaction	-0.173	0.339	.610	0.550	0.352	.118	-0.310	0.401	.439
Direct effect PE method -> satisfaction	0.058	0.352	.869	0.958	0.440	.029*	-0.194	0.420	.645
PE method -> perceived control	-0.600	0.220	.006*	-0.700	0.305	.022*	-0.565	0.288	.050
Perceived control -> satisfaction	0.386	0.188	.040*	0.583	0.346	.092	0.206	0.253	.415
Indirect effect PE method -> satisfaction	-0.231	0.139	.096	-0.408	0.247	.099	-0.116	0.162	.472

Table E1: Regression table of the mediation model for the effects on recommendation satisfaction in experiment 1, including stratified models with both levels of domain knowledge.

Experiment 2

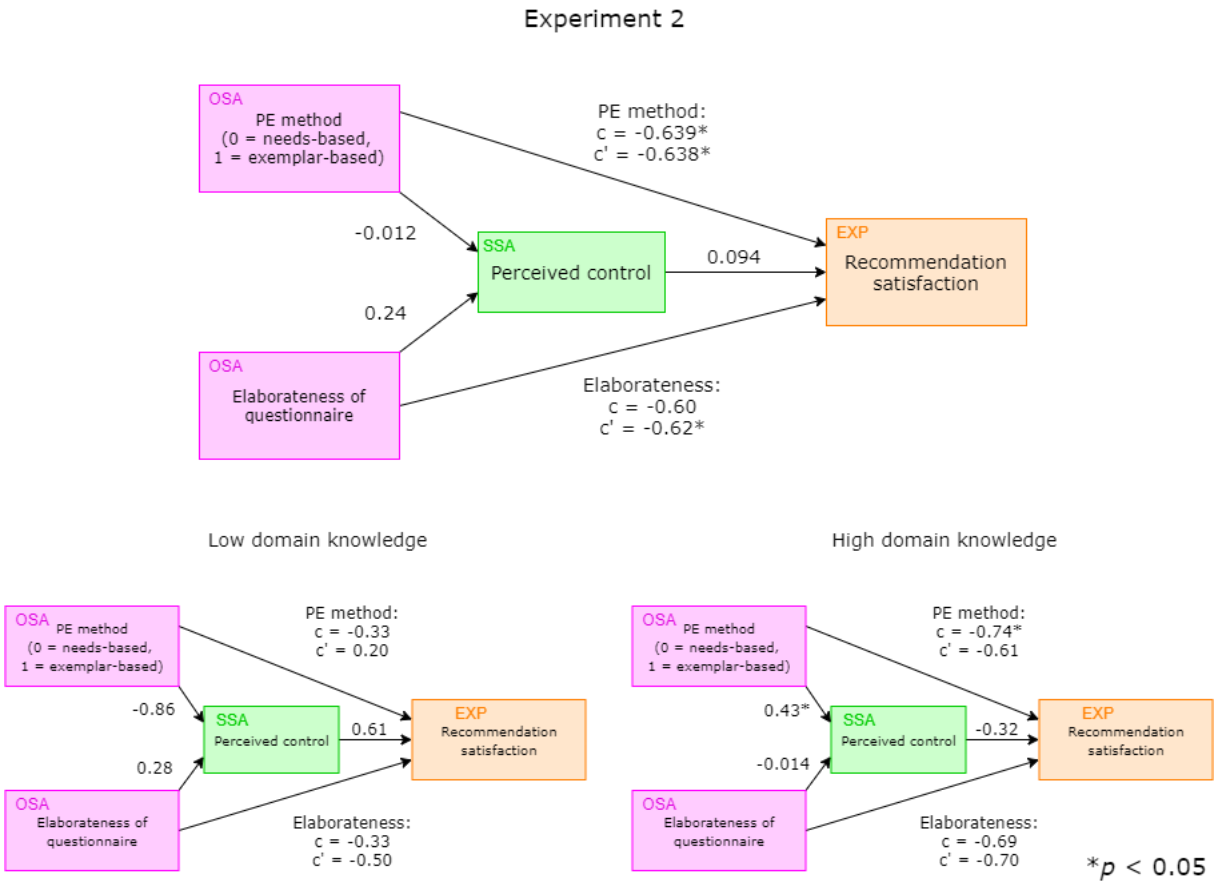

Figure E2: Results of a mediation analysis of experiment 2 (top), and the mediation analyses at each level of domain knowledge (bottom). For each effect, the regression estimate is displayed. An asterisk denotes a statistically significant effect.

	Full sample			Low domain knowledge			High domain knowledge		
Effect	*Estimate*	*SE*	*p*	*Estimate*	*SE*	*p*	*Estimate*	*SE*	*p*
Total effect PE method -> satisfaction	-0.639	0.307	.037*	-0.329	0.473	.487	-0.744	0.350	.033*
Direct effect PE method -> satisfaction	-0.638	0.311	.040*	0.198	0.591	.737	-0.606	0.337	.072
Total effect elaborateness -> satisfaction	-0.599	0.307	.051	-0.329	0.486	.499	-0.692	0.371	.062
Direct effect elaborateness -> satisfaction	-0.622	0.301	.039*	-0.503	0.484	.299	-0.696	0.377	.065
PE method -> perceived control	-0.012	0.215	.956	-0.859	0.537	.110	0.430	0.202	.034*
Elaborateness -> perceived control	0.244	0.216	.259	0.284	0.544	.602	-0.014	0.212	.947
Perceived control -> satisfaction	0.094	0.184	.609	0.614	0.317	.053	-0.320	0.319	.316
Indirect effect PE method -> perceived control	-0.001	0.044	.980	-0.527	0.609	.387	-0.137	0.157	.383
Indirect effect elaborateness -> perceived control	0.023	0.061	.704	0.174	0.392	.657	0.004	0.100	.964

Table E2: Regression table of the mediation model for the effects on recommendation satisfaction in experiment 2, including stratified models with both levels of domain knowledge.