Graduation Thesis for the Master of Human-Technology Interaction

Adaptive Advice

adapting a recommender system for energy-saving behaviors to personal differences in decision-making

Bart Knijnenburg

Table of contents

Summary

A short overview of the project

This thesis describes the graduation project of Bart Knijnenburg for the Master in Human-Technology Interaction at Eindhoven University of Technology.

The project is about recommender systems: technology (usually in the form of a computer program) that helps people in complex choice situations. Using a user-centered approach, we develop and test several potential improvements to recommender systems. Specifically, we argue that decision-makers differ on several personal characteristics that influence their decision-making strategy. In order to accommodate for these differences in decision-making, we let go of the one-size-fits-all approach of present-day recommender systems, and investigate the benefits of an adaptive interface. The actual benefits of this paradigm shift are tested using an online recommender system for energy-saving measures.

The project considers two personal characteristics, domain knowledge and choice goals (in the energy-saving use-case: ecological commitment). Our first experiment confirms that people with different levels on these characteristics prefer different recommender system interfaces. Specifically, since the choice goal of committed individuals is to save the environment, they want to be presented with an 'environmental benefits frame' that displays their decision in terms of environmental benefits, while less-committed individuals want to be presented with a 'personal benefits frame' that displays their decision in terms of personal (monetary) benefits. Furthermore, since novices easily experience information overload and evaluate choice options in a holistic fashion, they want a restricted amount of information and a case-based preference elicitation method, while experts, who are more familiar with the attributes of the decision domain and who use highly detailed information to make better decisions, want highly details information and an attribute-based preference elicitation method.

Whereas domain knowledge and commitment can be measured using questionnaires, such measurement is impractical in real-life implementations of recommender systems. In order to unobtrusively measure these user characteristics, we determine relations between the characteristics and observed differences in process data (clicking behavior). Using these process data relations, we define rules for a truly adaptive system: one that measures the users' domain knowledge and commitment on the fly based on the clicks in the interface, and uses it to update a 'user model'. Based on the values of this 'user model', the system then changes the interface during the interaction.

Several versions of the adaptive system are tested against a 'static' baseline system. Specifically, besides the baseline, three adaptive versions are considered: a system that adapts the interface without any explanations; a system that uses a generic pop-up to explain the adaptive behavior when it occurs; and a system that explains the adaptive behavior using a humanlike agent.

Our second experiment confirms the added benefit of our adaptive system: the system that employs generic explanations is found to provide better personal help to the user, and is perceived as more satisfying and useful than the static system. The experiment also confirms that explanations are necessary for adaptiveness to succeed: The system without explanations is found to be more confusing, and is perceived as less satisfying and useful than the static system. Contrary to our expectations, users of the agent-based system do not accept or understand the adaptiveness better than users of the other adaptive systems. Contrary to the system with generic explanations, the system with agent-based explanations is not perceived to be more satisfying or useful than the static system.

Although the results confirm the benefits of (adequately explained) adaptive recommender systems, these results are moderated by the type of preference elicitation that is used during the interaction. Specifically, the adaptiveness is rated more favorably when users make more use of the attribute-based preference elicitation. Furthermore, the results of our adaptive system may be attenuated due to the process-data measurement of user characteristics, which is far from perfect.

Taking a broad perspective, the project succeeds in combining fundamental principles of decision theory, psychology and interaction design to advance the field of recommender systems. User-testing of these advances ensures a user-focused approach. We contend that a multi-disciplinary approach can improve recommender systems and increase user satisfaction.

Introduction

This chapter briefly introduces the thesis, and gives an overview of the social context and the general problems that will be addressed in the research. It will argue that recommender systems should not retain their 'one-size-fits-all' approach, but instead adapt to the users' personal characteristics, specifically to users' level of 'domain knowledge' and their 'choice goals'.

The chapter will also introduce the case of energy-saving measures to demonstrate a typical case of personal differences in domain knowledge and choice goals. The remainder of this thesis uses this case to test the feasibility and the effects of adaptiveness.

Recommendation systems

Help me choose

Online recommendations: sell, advise, inform

Making choices is often a cognitively tasking endeavor, especially in our modern society where there are often too many options that are all attractive. For example, over three hundred types of digital cameras are being sold today, each of which can be described on at least thirty main attributes like resolution, zoom, storage media and battery life. In cases like this, the choice process can be typified as being complex, mind-boggling or downright annoying. Scholars in consumer behavior and marketing managers therefore acknowledge the potential benefits of recommender systems; computer programs that help their users make complex choices (Alba, et al., 1997).

Haubl and Trifts (2000) were arguably the first to systematically analyze the effects these recommender systems have on the decision maker. In their experiments, they used a system that provided recommendations based on attribute weights, and that displayed the products and attributes in a matrix that could be sorted by any attribute. Haubl and Trifts found that, compared to a static list of products, such a system significantly increased both the objective and subjective quality of the choices their participants made.

Outside academic life, merchandize sales systems have taken a central place in the field of retail, providing information and advising about choices. The advent of the Internet has brought online shopping sites like Amazon and Buy.com into the homes of millions of Internet-users. Most of the research on recommender systems, however, has focused on the technical aspects of providing recommendations and much less on user-related aspects (Xiao & Benbasat, 2007).

In this thesis, we claim that user-focused research is critical for the adoption of recommender systems, and we investigate the merit of several potential user-focused improvements.

Utility theory

Although a wide variety of recommender systems exists, their operation is usually based on the notion that a choice domain has certain universal attributes, and that choice options can be described in terms of the values of each of the attributes, like resolution, zoom, storage media and battery life in the case of digital cameras (Bettman, Luce, & Payne, 1998). An important distinction that can be made in the proliferation of recommender systems is whether the system uses a non-compensatory strategy for recommendations or a compensatory one (Guttman & Maes, 1998). A non-compensatory strategy defines rigid cut-off values for each of the attributes (e.g. a resolution of *at least* 8 megapixels), and only recommends those products that adhere to each of the cut-off values. In contrast, a compensatory strategy allows trade-offs between attributes, e.g. compromising on one

attribute (e.g. lower resolution) in return for a better value on another attribute (e.g. longer battery life).

Many e-commerce websites recommend products using a non-compensatory approach by allowing the user to restrict the set of possible choices by providing upper or lower limits on certain attributes (Xiao & Benbasat, 2007). For instance, when you are shopping for a laptop on HP.com, you can refine your search by choosing from distinct categories, like a certain price, weight, or screen size range.

This way of refining your search, which is called 'Elimination By Aspects' (Bettman et al., 1998) is very useful if want to narrow down your result set quickly, or if you have a clear optimal point on all the attributes and if you are unwilling to compromise on any of these points. If you have no clear idea about some of the attributes, or if you do not mind compromising on one attribute to get a much better value on another, the non-compensatory method is less suitable, as you have to change your restrictions several times to find the best choice option.

In such cases, compensatory strategies are more suitable. MAUT, or Multi-Attribute Utility Theory, is the best-known implementation of compensatory choice. In its simplest form, the MAUT method lets the user assign weights to each of the attributes. It also assigns a value to each attribute level of each product, multiplies these values with the user-assigned weights, and sums these to get the utility of the product for this user. The MAUT method assumes that an option with the highest utility for a certain user is preferred over the other options.

In our use case of energy-saving measures (which will be presented shortly), it is more natural to use a compensatory strategy than to define hard cut-off points for each attribute. Because of this, the current thesis addresses MAUT-based recommender systems only.

The 'one-size-fits-all' approach of present-day web shops

Although Haubl and Trifts (2000) proved the benefits of recommender systems, almost none of today's web shops apply any recommender system features beyond the non-compensatory 'drill-down'. Since Haubl and Trifts' paper, many other researchers have investigated the effects of recommender systems on the choices people make (for a review, see Xiao & Benbasat, 2007[1]), but have failed to bridge the gap between the theoretical advantage and the practical applicability of such recommender systems. A good way to find possible causes for the gap is to look at the problems that occur in online shopping.

Cox and Dale (2001) and the COGITO group (Andersen, Hansen, & Andersen, 2001) investigated the needs that customers in an online shopping environment have[2]. In the light

[1] As the focus of this thesis is on the cognitive and decision-theoretic aspects of recommender system use, the literature reviewed here is mainly experimental in nature. We acknowledge that there is also a vast body of survey research done on online shopping attitudes and behavior (for a review, see (Li & Zhang, 2002; Cheung, Chan, & Limayem, 2005)).

[2] Not surprisingly, a study determining the customer needs in the domain of energy advice, performed by Darby (2003), provided very similar needs.

of this thesis, their most important implication is that customers do not only differ on what product they want, but also on the way they want to be treated by a sales representative. In online retail, the first order of diversity is not a problem. In contrary, most web shops offer an order of magnitude more options that a normal shop. The second order of diversity, however, is usually not available in web shops.

Many web shops are mostly tuned to the more advanced end of the customer spectrum: people with a lot of 'domain knowledge'. These 'experts' have intimate knowledge of product attributes and can single-handedly search for the best product to match their demand. Novices, however, are often scared away by the myriad of options, the technical descriptions and the inability to express their preference with a simple question when shopping online (Alba, et al., 1997). Take for example electronics web shops, where the number of complex features can become puzzling, especially to the novice customer (Wang & Benbasat, 2007).

A traditional store is better suited for these novices, since the sales representative can act as a human 'interface' between their vague but practical demands and the complex and numerous products offered. Moreover, the traditional store can offer this service to novice buyers and provide quick and detailed information to advanced buyers at the same time.

Furthermore, sales assistants in these traditional stores can adapt to the customer's personal goals(Aberg & Shahmehri, 2000). By establishing the choice goal early in the conversation, they are able to frame their questions and recommendations in a way that matches this choice goal, making it easier to comprehend and choose. Online, however, information about the choice goal is lost, and all customer types have to browse the same set of all available products. As these online systems do not take choice goals into consideration, most web shops are designed for the 'typical' customer, and customers with a divergent choice goal suffer from a lower level of service.

The solution: adaptiveness

If a web shop could offer a diversity of service similar to a traditional store, it could extend its clientage to novices and customers with a divergent choice goal, thereby increasing sales. A similar remark was made by Spiekerman and Paraschiv (2002), who argue that "user interfaces proposed by marketers should not continue to follow the rule of 'one-size-fits-all' types of interaction as we can mostly observe it on the Net today. Instead, they should try to exploit consumers' expected involvement with the product as well as the perceived level and nature of perceived purchase risk[3]." (p. 281)

Letting go of the 'one-size-fits-all' solution, this thesis proposes to implement *adaptiveness* as a way to tailor the system to the users' needs. Specifically, the interface of a recommender system should be tailored to the level of customer *domain knowledge (or expertise)*. Furthermore, giving customers personal attention requires that the system understands the

[3] Perceived risk is commonly seen as a result of a certain level of expertise. Specifically, the lower the expertise, the higher the perceived risk of potentially making the wrong choice.

customers' *choice goal* (Xiao & Benbasat, 2007; Spiekermann & Paraschiv, 2002; Maes, Guttman, & Moukas, 1999). This thesis will introduce the idea of adaptation to domain knowledge and choice goals. Before going into details on these two types of adaptiveness, the use case of energy-saving measures will be introduced.

Use case

Energy-saving measures

Benefits of this use case

While this thesis endeavors to apply its implications to recommender systems in general, the experiments take energy-saving measures as a specific use-case. An online recommender system is developed that helps people choose which energy-saving measures to implement. The energy-saving use-case is a convenient one because:

- The case of saving energy is highly relevant in today's society, and we believe that recommender systems could provide an important contribution to the global reduction of energy-use[4]. Quantifying energy-saving measures in terms of a set of universal attributes is in itself a valuable undertaking.
- It is natural to make multiple choices in this context (in contrast to, for instance, buying a new computer), which produces more interesting data.
- Brand loyalty and other unquantifiable preference influencers are less prominent.
- We will not need to 'sell' our customers anything; the situation can be implemented in a highly realistic way without having to set up a sales infrastructure.
- People have different levels of domain knowledge about energy-saving measures, and this topic has been thoroughly researched.
- People have different fundamental goals in saving energy; some do it for the environment, while others do it to save money. There is a substantial body of research on this topic as well.

These points make energy-saving measures a convenient and rich case for this thesis.

Many websites exist that provide energy-saving advice, and some of these sites are able to tailor the measures to the users' living conditions. However, to our best knowledge, there exists no recommender system for energy-saving measures that provides recommendations based on the users' stated preferences. The study that is arguably closest to a recommender system for energy saving behavior was performed by Farsi (2008), in which he constructed a multi-attribute utility model to analyze people's willingness-to-pay for several energy saving

[4] Quite a lot of research has been done on attitudes towards energy-saving behavior (Kaiser, Wölfing, & Fuhrer, 1999). However, as Van Raaij and Verhallen (1983) note: "Through recommendations, information, prompts, and information about the energy costs of certain behaviors we may change behavior directly without changing attitudes first." (p. 60)

measures. However, beyond general policy recommendations, Farsi does not use the results of his analysis to provide energy saving recommendations.

An adaptive recommender system

For energy-saving behaviors

This thesis will address the opportunities for adaptiveness in a recommender system for energy-saving measures. Ecological knowledge (domain knowledge) and ecological commitment (choice goals) are singled out as personal characteristics that influence choice behavior. It hypothesizes that these characteristics are therefore possible subjects for adaptation, and that correct adaptation will likely result in a higher satisfaction and better choices. Two experiments will be described that test this hypothesis.

Theory and existing research

This chapter explains the idea of adaptive recommender systems, and argues that personal differences in choice behavior could guide such an adaptive approach. The chapter introduces two personal characteristics of recommender system users that may influence their choice behavior: domain knowledge and choice goals. It argues that different interfaces may be optimal for these different types of users, and that recommender systems therefore need to adapt their interface to these differences in order to increase satisfaction.

Furthermore, the chapter introduces the idea of a human-like agent to explicitly and implicitly explain the adaptiveness of the system.

Finally, the chapter will summarize the thesis by presenting a central argument that asserts that – compared to a 'traditional' recommender system – an adaptive recommender system with agent-based explanations will have positive effects on the satisfaction and the choices made by individuals using the system.

Adaptiveness

Lessons learned from existing research

What is adaptiveness?

The term 'adaptiveness' describes a wide array of practices in computer science research. Some disambiguation is required. 'Static' systems provide no adaptiveness, and look and behave broadly the same for all users. 'Customizable' systems are more adaptive; users can explicitly change the system to their liking (Höök, 2000). 'Tailored' interfaces automatically adjust to the user, but do this only once at the beginning of the interaction (Höök, 2000). An example of tailored systems is a website that detects the browser language and presents its content in that language too. An 'adaptive' system changes the interface on the fly (Jameson, 2002). It analyses the interaction, constructs a user model, and then updates the interface to match this model. Adaptive systems do not need initial knowledge about the user, and are able to evolve with the user if the users' preferred interface changes over time.

Although adaptive systems do not need initial knowledge about each *specific* user, they do need to have a preconceived (and often pretested) idea of what *types* of users there are, and what interface is preferred by each user type. This preexisting knowledge can however also be circumvented by adding an extra layer of adaptiveness that uses the outcome of the interaction (sales records, satisfaction ratings) to optimize the rules that govern the adaptiveness (e.g. Hauser, Urban, Liberali, & Braun, 2009). Such a 'multi-stage adaptive' system is however often too complex and unmanageable for real life usage, and is therefore beyond the scope of this thesis.

Adaptiveness in recommender systems

Adaptiveness can increase the similarity between the recommender system and the user, and this may lead to higher decision quality, lower decision effort, and higher levels of trust, satisfaction, and perceived usefulness (Xiao & Benbasat, 2007). Xiao and Benbasat note that it is therefore interesting to investigate whether adaptive recommender systems are more likely to be used.

Lu (1999) implemented a prototype of a fairly sophisticated adaptive web shop for buying toys. The system used if-then rules to expand user queries based on a user model. However, the Lu web shop was based on search, browsing and selection, and did not include a recommendation-module. Besides that, the user model was specified manually by the user. To our knowledge, Lu did also not test her system with real users.

A cautionary remark on adaptiveness

Hauser et al. (2009) recently claimed a rather substantial success in 'morphing' the information presentation of an online website selling broadband internet subscriptions for BT Group. Their system includes two levels of adaptiveness.

As a first level of adaptiveness, their system categorizes users on four dichotomous dimensions (which they call 'cognitive-styles') based on click stream data, and dynamically morphs three dichotomous interface aspects of the website according to these cognitive-styles. These interface aspects are graphical versus verbal presentation of quantitative data, few versus many products shown, and little versus much information displayed about each product. All three aspects are closely related to information load. As the second level of adaptiveness, the rules for morphing interface aspects based on cognitive-styles are optimized using a sales-optimizing criterion.

In other words, the system implemented by Hauser et al. is a 'multi-stage adaptive' system, in which even the rules for adapting the interface are themselves adaptively changed based on a higher criterion. Such a system runs the risk of 'over-automation' which causes the model to settle in trivial local optima. For example, the second level of adaptiveness may conclude that all cognitive styles benefit from the same interface aspects, which basically reduces the system to an optimized but static system.

It seems that this is exactly the case for Hauser et al.'s system. Although they report a substantive 20% increase in sales due to their system, a careful reader of their paper may notice that the bulk of this effect (13%, or about 63% of the 20% increase) is caused by simply finding the best *average* interface (the best combination, on average, of the three dichotomous interface aspects) *without* any user-based adaptiveness. Only a 7% increase in sales is therefore truly attributable to the sophisticated adaptiveness.

Furthermore, this 7% increase in sales is only a theoretical value, which can only be achieved in the theoretical case that the system can predict the cognitive styles perfectly. The actual system tries to predict the cognitive-styles based on process data, and this prediction is not without error. The error-prone click stream data provided only a further revenue increase of 0.6% instead of 7%.

This reinterpretation of Hauser et al.'s data makes two important points about adaptation. First of all, adaptation can easily be confused with optimization, the latter being a method to find the best *average* interface for the user population. Adaptation goes beyond this to provide a customized interface for a specific part of the user population, which is believed to increase the usability even further. Therefore, in order to prove the usefulness of adaptiveness, one should test the adaptive system against the *optimal* instead of a *random* combination of interface aspects. Furthermore, Hauser et al. may have 'over-automated' their system's adaptiveness, as the sophisticated click stream-based adaptiveness only accounts for a 0.6% increase in sales. Although detailed click stream data goes a long way in predicting purchase behavior (e.g. Van Den Poel & Buckinx, 2005), adapting the website to increase this purchase behavior is not trivial. Specifically, human interpretation of click stream data may provide larger usability gains than an automatic optimization process Langerwerf (2009).

The 'cognitive styles' defined by Hauser et al. are arbitrary dimensions that are not based on any scientific theory, but are defined solely as the result of the automatic optimization process. In contrast to this approach, we investigate the use of 'static' user types that are based on

individual differences in choice behavior as a basis of adaptiveness. Specifically, we use fundamental principles in decision-theory to reason that choice-behavioral differences between people with different levels of domain knowledge and with different choice goals are an adequate subject for adaptive changes in the recommender system interface.

Adapting to domain knowledge

In our case, ecological knowledge

Domain knowledge in general

'Domain knowledge' or 'expertise' can be described as a body of declarative and procedural knowledge about a certain domain. In choice situations, however, the concept can be further restricted to the part of this knowledge that is instrumental (or even required) to make adequate decisions. This includes knowledge about the attributes, their values, and the implications of certain values on product quality, as well as common trade-offs in making choices in the current domain.

Individual differences in domain knowledge have been thoroughly researched in the domain of decision making and consumer behavior (Alba & Hutchinson, 1987), but also in the specific field of online retail(Cheung, Chan, & Limayem, 2005; Li & Zhang, 2002). The most important issue concerning domain knowledge is that novices have less knowledge about the attributes of the choice options (Xiao & Benbasat, 2007). Consequently, novices have a harder time constructing their preference, as noted by Guttman (1998): "It is difficult to accurately express preferences for complex products, especially the first time a shopper is confronted with product features not considered before." (p. 36) Consequently, experts usually perceive a lower level of risk regarding the choice situation (Spiekermann & Paraschiv, 2002) while novices are usually more overwhelmed by the number of alternatives presented (Schwartz & Clore, 1988), because they typically lack an ideal preference point (Chernev, 2003).

This chain of effects has been researched by Coupey, Irwin and Payne (1998), who, in a series of experiments, found that when users had low product category familiarity, they made more preference reversals in choice versus matching tasks. Such a task-effect, they reason, is due to the fact that novices tend to focus on the most prominent attribute of the choice options in choice tasks, but not in matching tasks. This, in effect, is caused by the difficulty of constructing a preference in terms of attributes when dealing with unfamiliar product categories.

Likewise, in a review of literature on the effects of consumer expertise, Alba and Hutchinson (1987) find that experts are better at attribute weighting, focus on relevant attributes (instead of 'easy' or prominent attributes like opinions, prices and brands), and refrain from unfounded selective simplification.

These fundamental issues also hold for ecological knowledge, the energy-related variant of domain knowledge. As Parnell and Popovic Larsen (2005) note: "the expert is able to draw on an understanding of invisible processes and appropriate terminology to conceptualize energy, its use, and its conservation. Everyday householders, on the other hand, although it is possible that they too will understand expert vocabulary, will not necessarily find meaning in these concepts within the context of their everyday lives." (p. 796) Similarly, Darby (2003) finds that consumers in her study on energy advice "displayed differing levels of resources, confidence and ability to learn with and without guidance." (p. 1222) Consequently, she claims that "All householders bring their experience with them when they seek out or interpret advice and information, and it is a crucial part of the adviser's task to understand something of that experience and to build on it." (p. 1225)

Concerning recommender systems, these differences in choice task perception and behavior warrant the potentially successful application of adaptiveness to the level of domain knowledge of the user.

Measuring domain knowledge

Successful application of adaptiveness to domain knowledge requires a robust and consistent way to measure this user characteristic. An attempt to measure domain knowledge has resulted in a robust 5-item measurement scale developed by Flynn and Goldsmith (1999). In the case of energy-saving measures, the level of domain knowledge can be measured in terms of ecological knowledge, which for instance has been described in Kaiser, Wölfing and Fuhrer(1999). Kaiser et al. measure this concept by asking whether participants agree with certain ecological facts. This results in a general measure of ecological knowledge. In a choice situation, however, we want to measure the more specific knowledge needed to make trade-offs between choice options and attributes. Therefore, this study uses a more direct measure of the knowledge about and familiarity with energy-saving measures and their attributes specifically. Questions that measure this type of knowledge ask users to appraise their familiarity with energy-saving measures and attributes and their perceived ability to evaluate and compare measures. Specific questions can be found in Table 29 in Appendix F.

Domain knowledge can also be measured during the interaction. As experts are typically more capable to perform goal-directed search (Alba & Hutchinson, 1987), it can be expected that experts require fewer clicks per choice when using a recommender system. Also, experts will look more at detailed information, while novices prefer general information (Alba & Hutchinson, 1987). Compared to experts, novices are also more inclined to change their preferences repeatedly during the choice process, as they usually lack an ideal preference point (Chernev, 2003). A careful selection of process data predictors can be used to measure these differences.

In sum, this thesis hypothesizes that individuals with a higher level of ecological knowledge are more familiar with the presented attributes in our study, are more able to easily make

trade-offs between them, and possess on average more knowledge about the details of the available choice options. These aspects can be measured before and during the interaction.

Preference elicitation methods

Experts usually have well-articulated preferences (Bettman et al., 1998), which means that they generally have a lower perceived choice risk than novices (Spiekermann & Paraschiv, 2002). Preferences are often constructed on the fly (Bettman et al., 1998), using available examples as reference points (Pu & Kumar, 2004; Pu & Chen, 2005; Viappiani, Pu, & Faltings, 2007; Viappiani, Faltings, & Pu, 2006), and this is especially relevant for novices.

As experts are more familiar with the presented attributes, they have no problem assigning weights to these attributes, while novices on the other hand typically lack the knowledge to decide which attributes are important (Pu & Kumar, 2004; Pu & Chen, 2005; Alba & Hutchinson, 1987; Coupey, Irwin, & Payne, 1998; Xiao & Benbasat, 2007). Expert reasoning is often compensatory, meaning that experts are capable to allow higher values in one attribute to make up for lower values in another attribute. Novices often lack these compensatory reasoning skills and more often apply lexicographic decision heuristics (Bettman et al., 1998).

It seems that these disparate forms of reasoning could benefit from their own interface for eliciting the users' preference.

Attribute-based preference elicitation

The most extensively used preference elicitation method, attribute-weight selection, might be well-suited for individuals with adequate domain knowledge. In this method, users indicate the importance of each of the attributes with which the choice options are described.

Haubl and Trifts (2000) discovered that, compared to providing a static list of choice options, making a pre-selection of products based on user-assigned attribute weights reduced search effort, increased the quality and reduced the size of the consideration set (options seriously considered for the final choice), increased the objective and subjective quality of the choice, and increased the users' confidence in their choice. Furthermore, they discovered that, compared to the static list, a 'comparison matrix' which showed the products in its rows and the attributes in its columns and which could be sorted by a certain attribute increased the quality and reduced the size of the consideration set, and decreased post-choice switching. To sum up, Haubl and Trifts found that "the two interactive decision aids [pre-selection and comparison matrix] allow consumers to make much better decisions while expending substantially less effort" (pp. 17-18).

Olson and Widing (2002) did not find a similar increase in choice accuracy for a recommender system that ranked items according to user-assigned attribute weights. However, they did find an increase in satisfaction and a decrease in decision time. One might

infer that Olson and Widing also found an improvement for attribute-based preference elicitation.

Although the two studies cited above find a general increase in decision quality by using attribute-based preference elicitation, this thesis predicts that the specification of attribute weights works best when the users are familiar with these attributes, understand the value of each of them, and are capable of making trade-offs between them. This is usually the case for expert decision makers (Alba & Hutchinson, 1987; Coupey et al., 1998).
This argument can also be derived from two existing studies on recommender systems. Spiekerman (2001) argues that when customers know their preferences, it is best to ask for these preferences explicitly. This means that attribute-based preference elicitation would be best for experts, as they are more able to explicitly express their preferences in terms of attributes. Xiao and Benbasat (2007), on the other hand, argue that novice customers may not readily know how to express their preferences. These users may therefore need another preference elicitation method.

Case-based preference elicitation

An alternative preference elicitation method, case-based preference elicitation, might be more suitable for novices. Instead of assigning weights to attributes, the case-based preference elicitation approach allows users to evaluate entire choice options.
Guttman (1998) theorizes about such a system, in which customers can express a preference for one product over another, based on which the system can extract preferences. Guttman and Maes (1998) describe conjoint analysis as a way to infer the importance of specific attributes without asking consumers to rate these attributes explicitly. They note the merit of this preference elicitation method, but also note that a direct specification of product attribute values is faster and less susceptible to noise. Viappiani et al. (2006) also note that this process may be cognitively arduous.
A more convenient way to extract preferences from assessing examples is called 'Case-based Recommendation' (Smyth, 2007; Burke, Hammond, & Young, 1997; Jameson, 2002). This approach constructs preference models by analyzing the users' critique on certain examples. Within the case-based recommendation approach, a distinction can be made as how to critique the examples. Critique can be a simple comparison ("this is better/worse than the others"), or an expression of a trade-off in attributes ("this is good, but I want better X").
The comparison critiquing approach was first developed by McGinty and Smyth (2002). In their approach, the user can give either positive or negative feedback on choice options. The system then uses this feedback to change the preference weights. Trade-off critiquing involves picking an example and critiquing it using a trade-off. In other words, users evaluate the examples by showing what would make them better (Pu & Chen, 2005). However, as critiques

are often expressed in terms of attributes, this method of preference elicitation still requires a considerable amount of domain knowledge to be able to provide the critiques.

Smyth and McClave (2001) point out that, when predicting user preferences in case-based recommendation, one runs the risk of turning recommendations into a self-fulfilling prophecy. Users tend to agree with recommendations, which fortifies the currently held beliefs about the users' preferences (see also Pazzani & Billsus, 2002). Eventually, users will only be exposed to choice options that are in line with their initially held beliefs, and would not experience a wider range of alternatives. This is especially problematic when users construct their preferences on the fly.
To prevent this phenomenon, which Clark(2003, pp. 182-183) calls 'narrowing', it is advisable to separate recommendation from preference elicitation explicitly, for instance by having separate actions for *choosing* something (choice) and *liking* something (preference elicitation), and by performing the case-based elicitation on a more diverse set of choice options instead of the 'best' recommendations. Users then *indicate their preference* by evaluating choice options in a set providing a wide range of alternatives, while *choosing* options from a different set providing a narrow range of optimally fitting recommendations.

The former set with a wider variety of alternatives can be provided by what Viappiani et al. (2006; 2007) call the "look ahead principle": choice options that are selected based on this principle generally adhere to the current attribute weights, but differ to the point where one attribute gains in importance at the expense of another. In this way, the options show the possible outcomes of the compromises the user can make to the current recommendations (see also (McSherry, 2003)).
In an experiment extending Pu and Chen's previous trade-off critiquing system (Viappiani et al., 2006), they found that providing such look ahead examples primed participants to provide more critiques. This significantly increased their decision accuracy from 45% in the regular trade-off critiquing experiments to 80% (a later study showed an increase of 70% (Viappiani et al., 2007)).
Although researchers contend that case-based preference elicitation results in better decisions and higher satisfaction across the board (Viappiani et al., 2007), the current thesis predicts that experts with enough knowledge about the meaning of the attributes would prefer a direct assignment of attribute weights over this latent one. Novice decision-makers, however have less knowledge about the attributes and therefore tend to evaluate choice options in a more holistic fashion. The case-based preference elicitation method is therefore perfectly tailored to their needs, as it allows users to evaluate entire choice options. An additional advantage is that this method is conversational. The examples that the users critique are feedforward for their actions: the examples show what products would show up if a certain attribute was found to be more important. This feedforward helps users to develop their preference incrementally during the choice process, which is a strategy typically preferred by novices (Guttman, 1998).

Needs-based preference elicitation

A third preference elicitation method is 'needs-based preference elicitation' (Randall, Terwiesch, & Ulrich, 2007). In this method, the users indicate to what extent they do or do not have certain needs related to the product category. These needs are then translated into attribute weights by the system. This means that the system has to know what kind of needs could be of importance concerning the product category, and how a certain value on these needs can be translated into attribute weights.

Although we understand the merit of needs-based preference elicitation, the method is not generically applicable since one has to identify needs and relate them to attribute weights. This extra information makes the method also less compatible to attribute-based preference elicitation, whereas the outcomes of case-based preference elicitation can be translated into attribute weights without needing extra information.

Adapting preference elicitation method to user domain knowledge

We state the hypothesis that whereas case-based preference elicitation is probably better for novices, experts are better off using an attribute-based preference elicitation method[5].

A study on product customization by Randall et al. (2007) also indicates that matching the preference elicitation method to domain knowledge may be beneficial. Specifically, they find that experts are more satisfied with the system when they use a system with parameter-based preference elicitation (a non-compensatory, attribute based preference elicitation method), while novices are more satisfied with a system that uses needs-based preference elicitation (a compensatory method that lets the user assign weights to 'needs' that are linked to attribute values).

As a second hypothesis, we predict that a system which effectively *adapts* the preference elicitation method to the users' domain knowledge will enjoy higher levels of satisfaction and usability than a system without such an adaptive quality.

This hypothesis has been argued before[6]. Spiekerman and Paraschiv (2002) contend that current recommender systems fail to motivate user interaction because they limit communication with the user to product attributes and fail to adjust to the level of expertise a buyer brings into the purchase process. They propose a strategy to integrate different knowledge levels in the system by offering a different interface for experts and novices.

Provisions concerning preference elicitation adaptation

Two provisions have to be made concerning the adaptiveness hypothesis. First of all, adapting the preference elicitation method to the level of domain knowledge is especially useful for novice users, who usually have unstable preferences. Reviewing the literature on

[5] Please note that a preference elicitation method is not the same as an information processing method. In fact, novices are more likely to process information attribute-by-attribute (Bettman & Park, 1980).

[6] But this hypothesis has – to our best knowledge – never been explicitly tested.

recommender systems, Xiao and Benbasat (2007) concede that expert users of recommender systems are likely to have more stable preferences, and will therefore interact less intensively with the preference elicitation part of the interface. These users will therefore be less affected by the preference elicitation method of the recommender system.
Moreover, it is possible that an automatic adaptation of the elicitation method may be perceived as a loss of control. The loss of control may result in decreased trust, satisfaction, and perception of usefulness (Xiao & Benbasat, 2007).

Amount of detail in information presentation

Besides a difference in the way experts and novices construct and explicate their preference, they also differ in the depth of information processing and preferred cognitive load during the choice task. Novices, make suboptimal choices in the face of information overload. Specifically, Bettman, Luce and Payne (1998) contend that novices and experts use different choice heuristics. Specifically, they state that shortcuts and heuristics are more readily taken when information load is high, and that in these cases choices will become suboptimal when the wrong shortcuts are taken. Consequently, too much information may reduce the decision quality of novices.
Experts on the other hand, are known to actively seek more detailed information that will help them make better informed choices. Experts are able to increase their depth of search because they have a better conceptual structure of the knowledge needed to make decisions (Alba & Hutchinson, 1987).
It thus seems that while experts prefer highly detailed information to support their decision making process, novices require low levels of information detail to make their decisions.

Adapting information detail to user domain knowledge

Information detail adaptation has been researched before. The PUSH system (Höök, 1998) adapts the presented information to the user using a stretchtext technique (paragraphs that can be expanded and collapsed). The system aspires to reduce information load without significantly increasing the number of expand actions, and is evaluated favorably as compared to a system that collapses all information initially.
Besides the PUSH system, the system developed by Hauser et al. (2009) (discussed before) incorporates adaptations that are related to information detail. Hauser et al., however, do not evaluate their system with real users.
This thesis hypothesizes that a system that adapts the amount of information detail to user domain knowledge will enjoy higher levels of satisfaction and usability than a system without such an adaptive quality.

Summary

To summarize the current section on "adapting to domain knowledge", experts and novices require disparate preference elicitation methods and amounts of detail in the provided information. The first main hypothesis of this thesis is therefore that it would be beneficial to measure the level of domain knowledge and adapt the recommender system to it. Specifically, experts should be presented with detailed information and an attribute-based preference elicitation interface, while novices should be presented with general information and a case-based preference elicitation interface.

Adapting to choice goals

In our case, ecological commitment

Choice goal categorization

Marketers find it convenient to categorize consumers in relation to their goals. On a large scale, this means that products can be tailored to a certain market segment, e.g. Diet Coke for people who are concerned about their weight versus Coke Zero for people who are concerned about healthy food in general. On a smaller scale, companies construct means-end chains (Gutman, 1982) that link their product features to the goals of their various customers, e.g. a car manufacturer promotes 'safety and economy' in its ads for mid-sized family cars and 'freedom and adventure' in their SUVs.

The fact that people have different goals in mind when making decisions is a vital characteristic to be considered in providing decision support through a recommender system. Spiekermann and Paraschiv (2002) contend that current recommender systems fail to learn about buyers' characteristics and suggest that future systems would have to understand their choice goals in order to make better recommendations. Spiekermann (2001) suggests that this can be done by having a learning algorithm identify the user's *reasons* for using the recommender system, and that these identified reasons can be used to make to adapt the interface of the recommender system to the user's choice goals. Guttman et al. (1998) predict that "matching the system's user interface with the consumer's manner of shopping will likely result in greater customer satisfaction." (p. 153)

As goals differ across people, adapting the system to these goals seems to be a viable approach. This thesis analyzes one adaptation to choice goals, namely the framing of presented information.

Ecological commitment

An in-depth interview with an 'energy-saving consultant' revealed that the main goal distinction in home energy saving is whether people save energy for personal financial reasons or to save the environment. Specifically, Individuals with low ecological commitment

have a financial goal in choosing energy-saving measures, while individuals with a high ecological commitment have an environmental goal. Similarly, Parnell and Popovic Larsen ((2005), see also (Stern, 2000)) make the distinction between individuals that are motivated by social altruism and individuals that pursue an individual benefit that outweighs the cost of the energy-saving measure. Although these goals are not mutually exclusive, this thesis hypothesizes that which of the two goals is most important depends on the ecological commitment of the person. Specifically, individuals with low ecological commitment may still implement energy-saving measures but do this primarily to save money, while for individuals with high ecological commitment saving the environment is the main goal of saving energy and financial benefits are a side issue (Dietz, Stern, & Guagnano, 1998).

Measures of ecological commitment are provided in a proliferation of studies on determinants of environmental attitudes and behavior (Dietz et al., 1998; Barr, Gilg, & Nicholas, 2005; Van Raaij & Verhallen, 1983). A distinctive property in these studies is whether ecological commitment is measured as an attitudinal construct like the New Environmental Paradigm (Dunlap, Van Liere, Mertig, & Jones, 2000), or behavioristically, as is done by Kaiser (1998) in his scale for General Ecological Behavior.

Considering the use of ecological commitment in a recommender system for energy-saving measures, ecological commitment can also be measured during the interaction. In a choice situation, individuals with a high ecological commitment will focus their search on saving energy instead of money and will pay special attention to the lifecycle environmental costs. Compared to less committed individuals, they will also choose to implement more measures that are more difficult or costly to perform, and it is likely that they already perform some of the measures they come across. Less committed users, on the other hand, will pay special attention to low-effort measures with a high return on investment.

In a study on causes of environmental behavior, Dietz et al. (1998) found that individuals' trade-offs between environment and economy predicts four out of five of their behavioral indicators. They also found that ecological commitment predicts behavioral intentions better than past behavior. We hypothesize that in a recommender system, this will show in the process data predictors of environmental commitment. Specifically, clicks indicating intentions towards energy saving will be more predictive than clicks indicating past behavior.

Information framing

Individuals with different choice goals look for different types of information. A typical way for sales assistants to help customers in their choice process is to 'frame' the provided information in a way that matches their choice goals.

Considering energy-related choices, individuals with different levels of ecological commitment can either have a personal or environmental goal in choosing energy-saving measures (Parnell & Popovic Larsen, 2005). Likewise, energy-saving measures themselves can be framed either in terms of environmental benefits way or in terms of personal benefits. The 'environmental

benefits frame' could describe measures in terms of saved kilowatt-hours, and put emphasis on the total environmental costs of the entire lifecycle (production, operation and disposal) of the measure. The 'personal benefits frame' could describe measures in terms of Euros saved, and put emphasis on the increase (or decrease) in comfort when implementing the measure.

Adapting information frame to user choice goal

In a qualitative research study about energy advice by Darby (2003, p. 1224), an energy adviser claimed that it would be useful to be able to record the customer's goal, so as to be able to shape the advise around this goal. Parnell and Popovic Larsen (2005) note that there are multiple motivations for environmentally responsible behavior. They contend that if a policy program acknowledges these multiple motivations, it will appeal to a broader range of individuals, as they each perceive the program in their own context.

Different choice goals thus require different representational frames. The second main hypothesis of this thesis is therefore that it would be beneficial to measure the users' choice goal and adapt the recommender system to it. Specifically, ecologically committed individuals should be presented with an 'environmental benefits frame', while less committed individuals should be presented with a 'personal benefits frame'.

Adaptiveness and agents

The perfect fit

Adaptation means confusion: the problem of adaptive systems

We predict that an adaptive recommender system produces better choices and higher satisfaction than a non-adaptive one. However, the dynamic nature of an adaptive system may cause the user interface to change significantly when the system makes adaptations while the user is using the system. Especially in an online environment, where websites usually have a static structure, such sudden changes may confuse the user (Pazzani & Billsus, 2002).

Adaptiveness has been thoroughly researched in general (Jameson, 2002; Höök, 2000) and to a lesser extend for recommender systems specifically (Pazzani & Billsus, 2002), and there are even some usability evaluations of adaptive systems (e.g. (Höök, 1998)). However, although understandability issues have been suggested (Pazzani & Billsus, 2002), they have only sporadically been researched (Olson & Widing II, 2002).

We hypothesize understandability issues in an adaptive system, but at the same time, we provide a possible solution to this problem in the form of a human-like agent that explains the adaptive behavior of the system.

Agents and recommender systems

The interaction concept of human-like agents has repeatedly been used in recommender systems. Bickmore and Cassell (2001) tested whether an agent that employed small talk to increase engagement, gains users' trust. They found that trust only increased with extravert participants, and that engagement only increased with participants that were inclined to initiate interactions.

Spiekerman (2001) included a human-like agent in an online recommender system for jackets and digital cameras. The agent was used for social facilitation, and to draw users' attention to certain parts of the interface.

Pazzani and Billsus (2002) also implemented a recommender system with a human-like agent. Their agent recommends papers to read based on previous choices. Pazzani and Billsus used a very pragmatic evaluation criterion for their agent (increased website traffic) and did not evaluate the user satisfaction or choice quality.

The COGITO project (Abbattista, Lops, Semeraro, Andersen, & Andersen, 2002) developed an adaptive online bookshop that uses a human-like agent with a natural language interface as its main interaction paradigm. A tree learning algorithm analyzed the users' interaction behavior related to each of the ten book categories in order to construct a user profile. A later extension to the COGITO system also analysed the conversation between the user and the agent to provide a finer grained adaptation(Semeraro, Andersen, Andersen, Gemmis, & Lops, 2008). The COGITO system, however, primarily provided a search facility and did not explicitly support the decision making process.

None of the aforementioned studies that included human-like agents explicitly tested their agent-based system against a system without the agent. However, several studies have found that certain properties of the appearance of a web shop agents – whether they were perceived as life-like, attractive or an expert – influenced user evaluations of their appropriateness and usefulness (Keeling, McGoldrick, Beatty, & Macaulay, 2004; Holzwarth, Janiszewski, & Neumann, 2006; McBreen & Jack, 2001).

Explanations and recommender systems

Several studies have investigated the effect of explanations (with or without the help of agents) in recommender systems research.

Liu and Benbasat(2005) implemented an agent-mediated live help system in a web shop environment. They found that users' perception of flow was increased when the live help employed text-to-speech interaction instead of a text-based chat, and that users' feelings of telepresence increased when the live help employed a 3D avatar. The study did not use autonomous agents, as the live help was operated by a sales representative. Liu and Benbasat also did not measure differences in purchase behavior. In a similar study using autonomous agents, Liu and Benbasat(forthcoming) found that human speech and 3D avatars increased

social presence, which in turn increased trust, perceived usefulness and enjoyment. These, in turn, increased the intention to use the system.

Likewise, Wang and Benbasat (2007) argue that "many [recommender systems] still lack adequate explanation facilities" (p. 218) and that "by making the performance of systems transparent to users, [explanation] can improve users' trust in the systems, facilitate the transfer of knowledge to users, and lead to more effective use of the systems and better product choices" (p. 218). In a series of experiments, they evaluate a recommender system that provides explanations for the preference elicitation procedure in terms of "how"-explanations (how a stated preference leads to a set of recommendations), "why"-explanations (why a certain preference elicitation question is asked) and "trade-off"-explanations (what the consequences are of a certain choice). They find that how- and why-explanations increase user belief in the benevolence of the system, and that additionally, how-explanations increase user belief in the competence of the system. Trade-off-explanations increase user belief in the integrity of the system.

Aberg and Shahmehri (2000) took a different approach. They implemented an online assistance functionality that is basically a chat session with a real human sales representative. Although they do not compare their system against a system without online assistance, they find that the human assistance helps users to gain confidence and reduce confusion.

Agents to explain

Although human-like agents have been used in choice situations before, and that explanation has been offered in existing research as a way to reduce the possible confusion that results from the dynamic nature of recommender systems in general, using an agent to explain the *adaptive qualities* of a recommender system is a new approach.

We hypothesize the benefits of such an agent to be twofold. First of all, the agent can explicitly explain the occurrence of an adaptation, by stating what has changed and why it changed. But more importantly, an agent implicitly explains the adaptive behavior by representing the autonomous behavior of the system. When an adaptation is made, the agent can explain that *it*, instead of the system, performed the change. The agent then appears to be an autonomous body that monitors the users' interaction, reasons about their domain knowledge and choice goals, and adjusts the system accordingly. In other words, its human-like appearance can be used as an instant metaphor for autonomy and intelligent adaptiveness.

This assertion is in line with our previous research (Knijnenburg & Willemsen, 2008). In a Wizard-of-Oz experiment testing the usability and interaction dynamics of agent-based interaction, we found that the human-like appearance of an agent entices users to heuristically reason about its functionality. Specifically, we found that users interacting with agent-based systems construct a use image of human-like intelligence, and thereby infer the presence of certain typically human-like capabilities like advanced language processing, context-awareness, inference of implicit information, connectedness of successive interaction, and the

ability to handle multiple requests at once (see also (Bickmore & Cassell, 2001)). This thesis predicts that adaptiveness is another quality that is inferred for agent-based systems, such that the occurrence of adaptiveness seems more straightforward and acceptable when the system has an agent in its interface.

Agents and expectancies

The 'intelligent' human-like agent, however, is subject to a new interaction paradigm that is notoriously hard to control (Keeling et al., 2004). Users infer the use image of an agent-based system from the way it 'looks' and 'talks', just like they would do when interacting with other human-beings (Cook & Salvendy, 1989). The fact that the 'system' is 'human' provides them *instantaneously* and *effortlessly* with expectations of human-like intelligence (Laurel, 1990). Furthermore, because of their instantaneous nature, these expectations cannot be traced back to the individual features of the agent's appearance; instead, the entire body of features is *integrated* (perceived as a whole) to construct a use image with a set of interrelated beliefs (Knijnenburg & Willemsen, 2008).

Knijnenburg and Willemsen (2008) show that serious usability breakdowns can happen when the system is not able to live up to the expectations of human-like intelligence. However, since the use image is integrated, it is inherently difficult to tweak the agent features in such a way that the user expectations exactly match the capabilities of the system. The resulting overestimation, therefore, is an inherent problem in agent-based interaction.

In the specific case of recommender systems, Guttman, Moukas and Maes (1998) introduce the idea of sales agent avatars, and contend that these anthropomorphized agents provide personal attention and a more engaging experience that resembles a real-world shopping environment, but that the technology behind the agents cannot yet meet their users' expectations. Moreover, using semi-structured interviews with 30 online shoppers evaluating human-like agents in online retailing, Keeling et al. (2004) found that there needs to be a match between expectations raised by the agent and the actual capabilities of the system.

This leads to the hypothesis that although an adaptive system may benefit from a human-like agent, it is possible that users may be confused by a mismatch between their expectations of the system and the actual capabilities of the system, or that they will be disappointed when they have to adjust their overestimated beliefs about the capabilities of the system. This may also explain why some e-commerce applications of human-like agents get a disappointingly negative evaluation (Andersen & Andersen, 2002; Bickmore & Cassell, 2001).

Summary

The reasoning above leads to the prediction that an adaptive recommender system can benefit from the introduction of a human-like agent that both implicitly represents and explicitly explains the adaptations of the system, but that it is possible that the agent will disappoint users that overestimate its capabilities.

A 'good' recommender system

How to measure recommender system performance?

Objective performance measures

It is common practice in recommender system research to measure performance 'objectively' in terms of the match between chosen items and the user's preference. For an investigation of preference elicitation methods, however, this type of measures is ineffective, because it assumes that the preference elicitation is without error.

Another common way to measure the objective performance of recommender system is to set an external goal of maximizing sales or profit. Such a performance measure may be useful for the owner of the recommender system, but does not necessarily give insight in the performance for its users. Moreover, higher sales may actually indicate a lower performance for users, since better recommendations lead to targeted sales.

Subjective performance measures

Employing subjective measures, one can measure user-related performance of recommender systems in a non-trivial way. Although measures like 'satisfaction', 'understandability' and 'perceived usefulness' are harder to interpret than sales figures and root mean square recommendation error, they nevertheless provide valuable insight in the customers' experience, especially when combined with an analysis of the decision-making processes employed by the customers.

Li and Zhang (2002) note that the attention in current Internet buying behavior research for consumer satisfaction is uncharacteristically low, considering the relative importance of this concept, especially when it comes to repeated use of the service. This situation has improved in recent years, but as Xiao and Benbasat (2007) note, the empirical research on recommender systems is divided in studies focusing on decision-making processes and outcomes, and studies focusing on users' subjective evaluation. This thesis tries to integrate these two streams of research, covering both decision process and system satisfaction as its subject of analysis. Based on an extensive literature review, Cheung et al. (2005) list the factors that may be subject to consumer satisfaction: convenience, ease of use, information quality, navigation, security, shopping aids, and usefulness. This thesis will evaluate the recommender systems employed in its experiments using a wide range subjective measures.

Central thesis argument

Adaptive recommender systems with agent-based explanations

Summarizing the arguments provided above, this thesis revolves around the following argument:

Compared to a 'traditional' recommender system – an adaptive recommender system with explanations will have positive effects on the satisfaction and the choices made by people using the system. Two potentially promising dimensions for adaptation are domain knowledge and choice goals, as these two dimensions are easily measured before and during the interaction, and require distinct and mutually exclusive interaction methods or information representations.

In the light of the energy-saving measures case, this argument can be framed as follows:

Individuals with extensive ecological knowledge should be presented with detailed information and an attribute-based preference elicitation interface, while individuals with little environmental knowledge should be presented with general information and a case-based preference elicitation interface.

Furthermore, as the goal of ecologically committed individuals is to save the environment, they should be presented with an 'environmental benefits frame' that puts environmental savings above effort and costs, while less committed individuals – who are more concerned with the financial gains of saving energy – should be presented with a 'personal benefits frame' that focuses on saving money while minimizing decreases in personal comfort.

In the light of adaptiveness, we add the following statements:

Adaptiveness can only succeed with adequate explanation facilities. Specifically, due to understandability issues, an adaptive recommender system without explanations will be perceived as less satisfying and less useful than a static recommender system. An adaptive recommender system with explanations will be more usable than a static recommender system.

Furthermore, agents are a natural means to explain adaptiveness, as adaptive behavior is implied by their human-like appearance. This means that agents may potentially increase the understandability and acceptance of the adaptive behavior of a recommender system. However, an agent's capabilities may be overestimated, and this may actually reduce the understandability of the system, as the user may expect the system to have capabilities that it actually does not have.

An adaptive recommender system

This chapter describes the workings of an adaptive recommender system for selecting energy-saving behaviors. The system is built in such a way that it can be used to test our central thesis argument in a series of experiments.
First, requirements for the system are derived based on the hypotheses to be tested. Consequently, the system is outlined to show how it meets the specified requirements. The adaptive behavior is then described in more detail. Finally, an experimental plan is described that shows how the two experiments conducted together test the validity of our central thesis argument.

Requirements

Designing a system to test the central thesis argument

In order to test our central thesis argument, we developed the Web Recommender System, an online recommender system that can monitor process data, incorporates the different requirements for different user types, and can change these features on the fly based on process data inferences. Specifically, the system has the following requirements:

- A MAUT-based recommendation strategy.
- Two different preference elicitation methods (case-based and attribute-based).
- Two different types of information about each choice option (general and detailed).
- The ability to sort on any attribute and display totals in either Euros or KWh (in order to present an environmental benefits frame or a personal benefits frame).
- Detailed registration of process data, cf. all clicks that the user makes. These data can eventually be used to measure the user characteristics (domain knowledge and commitment) on the fly.
- A user model that updates values of these characteristics based on process data rules, and the ability to change the aforementioned features when a threshold in the user model is crossed.
- The ability to explain the adaptations using either a generic explanation frame or a human-like agent.

The developed system includes 80 energy-saving measures (see Appendix A), which are described on 8 attributes (see Appendix B).

The system will be described below; refer to Appendix D for details on the evolution of the system, and to Appendix E for a technical description of the system.

Description of the system

How the Web Recommender System works

Overview of the system

The system consists of three parts (see Figure 1). The top part of the interface shows the preference elicitation. In this part, the user can set preference weights, either by positively/negatively evaluating example choice options (case-based preference elicitation), or by increasing/decreasing attribute weights directly (attribute-based preference elicitation).

The middle part shows recommended choice options, which are selected using MAUT on the attribute weights taken from the preference elicitation in the top part. Users can click on these recommended choice options to read more information about them, and to choose them (see Figure 2). A choice can mean that the user wants to implement this energy-saving measure ("I want to do this"), or that the user has already implemented it ("I am already doing this").

The bottom part contains two lists of chosen measures ("what I want to do", and "what I am doing already") and total amount of energy/cost savings.

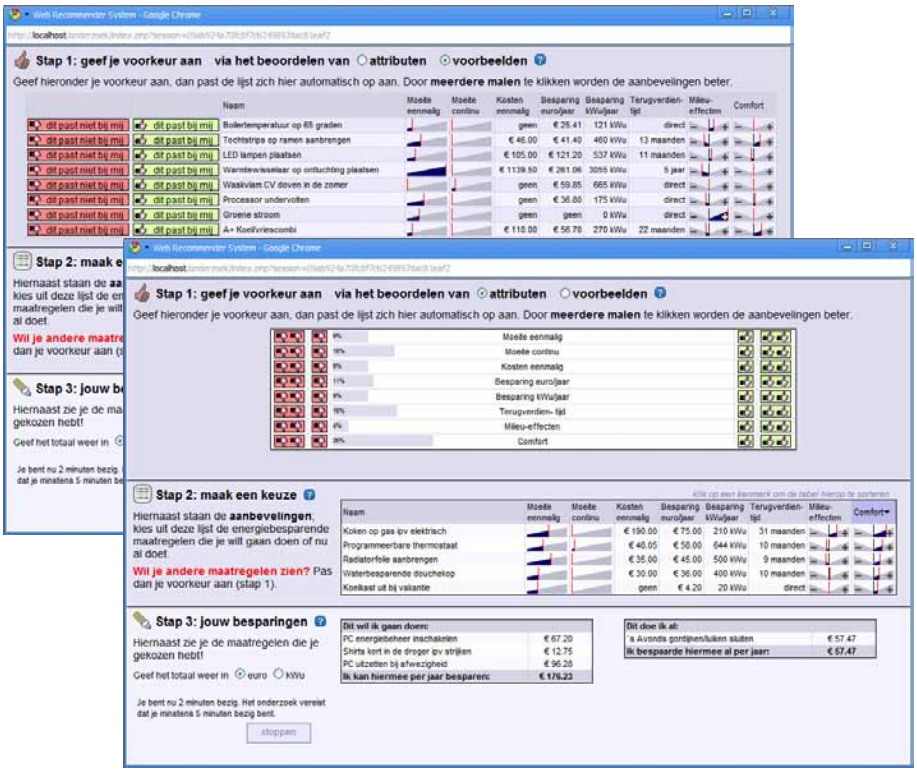


Figure 1: The Web Recommender System with case-based preference elicitation (upper screenshot) and with attribute-based preference elicitation (lower screenshot)

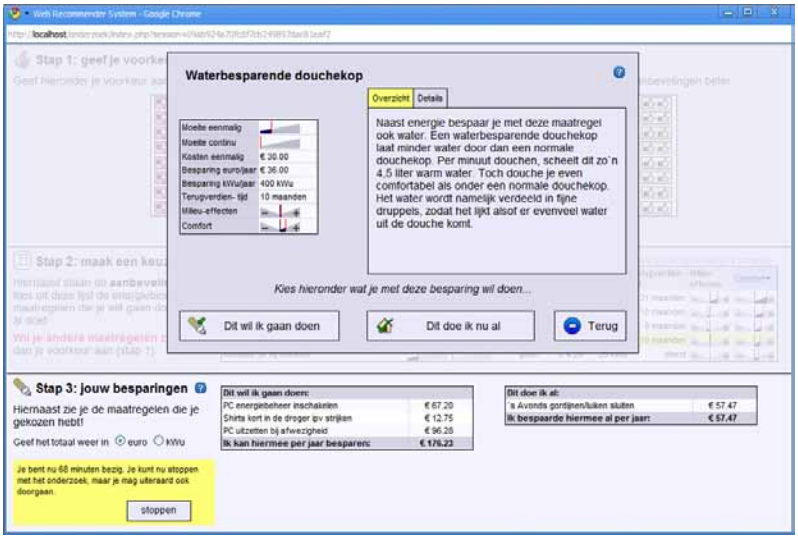


Figure 2: The Web Recommender System with a choice option selected

MAUT-based recommendations

As argued in the theory chapter, this thesis concerns recommender systems with compensatory strategies, and employs MAUT, or Multi-Attribute Utility Trade-off to select the

best choice options for the user. The MAUT method in our system computes the value of each choice option by assigning a value to each attribute value of the option, multiplying these values with the user-assigned weights, and summing them to compute the utility. The five options with the highest utility are displayed as recommendations in the middle part of the interface (see Figure 3).

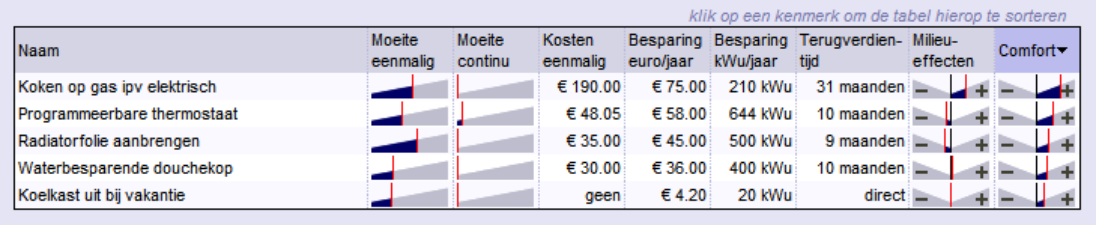


Figure 3: The recommendations that have the highest utility considering the user's preference weights

Preference elicitation methods

The system provides two preference elicitation methods. The attribute-based preference elicitation (see Figure 4) shows each attribute with buttons to increase or decrease their weight by one unit (one thumb up/down) or five units (two thumbs up/down). Since the attributes have different scales, a separate utility model calibration experiment was conducted to normalize their values (see Appendix C). These normalization weights ensure that one unit increase in attribute A has the same subjective impact as one unit increase in attribute B.

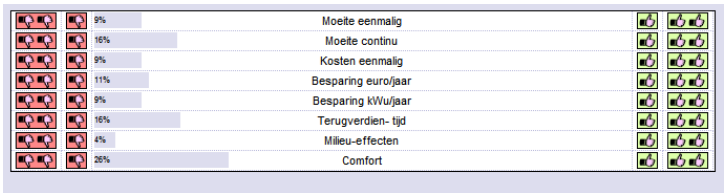


Figure 4: Attribute-based preference elicitation; the red and green buttons are used to decrease and increase the weight of the attributes

For the implementation of a case-based preference elicitation method, we combine the intuitive appeal of the comparison critiquing (McGinty & Smyth, 2002) with the 'look ahead' principle (Viappiani et al., 2006; 2007). Specifically, we provide a separate list of trade-off recommendations (see Figure 5), where each recommendation is selected using MAUT on the current preference weights with one modification: for each of the trade-off recommendations one attribute has been made more important (cf. its weight is increased by five units). This way, each of the trade-off recommendations shows a choice option that would be recommended if the 'important' attribute would have a higher weight (in other words, the system 'looks ahead').

Consequently, a positive (negative) evaluation of a trade-off recommendation can be treated similar to the increase (decrease) of the weight of its 'important' attribute. The system interprets these evaluations as increases or decreases of two units.

Figure 5: Case-based preference elicitation; every choice option represents the effect of increasing one of the attributes, the red and green buttons can be used to increase or decrease the weight of that attribute

The main merit of the two preference elicitation methods as described above is the fact that they are conceptually equal; they both update an underlying model of attribute weights that is used in MAUT-based recommendation. In other words, the recommendation algorithm is the same for both variants; only the elicitation method (in other words, the presentation to the user) differs.

Information types

The system provides two types of information: general and detailed (see Figure 6). The general information is usually a short description of the choice option in plain language. The detailed information provides deeper understanding of the choice option, and technical language and calculations are not avoided.

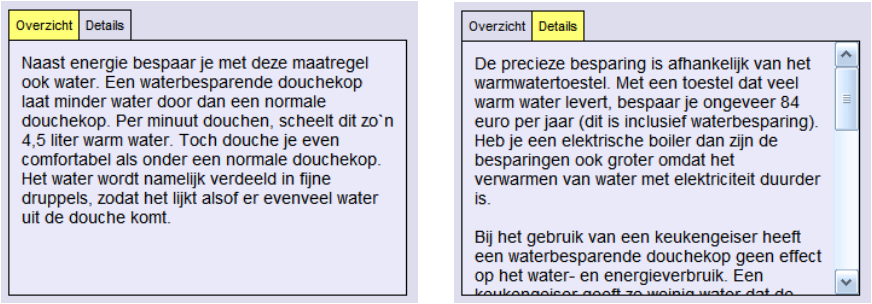
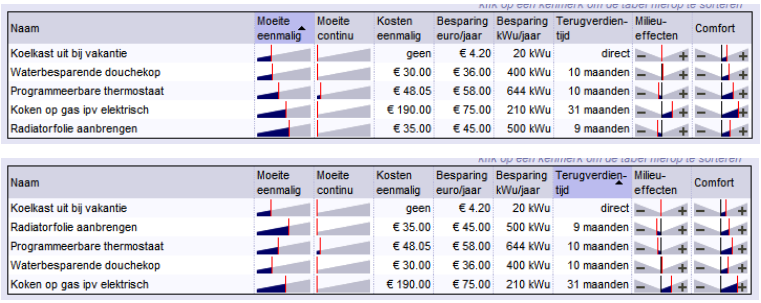


Figure 6: Information about a choice option; the system includes both general and detailed information

Sort and display totals

The list of recommendations can be sorted on any attribute (see Figure 7).

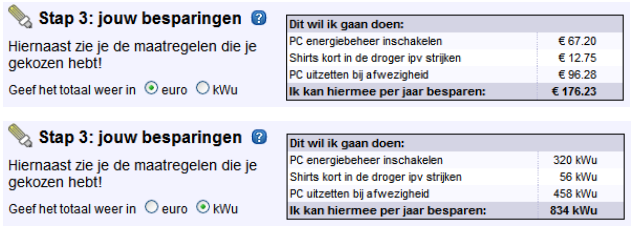


Naam	Moeite eenmalig	Moeite continu	Kosten eenmalig	Besparing euro/jaar	Besparing kWu/jaar	Terugverdien-tijd	Milieu-effecten	Comfort
Koelkast uit bij vakantie			geen	€ 4.20	20 kWu	direct		
Waterbesparende douchekop			€ 30.00	€ 36.00	400 kWu	10 maanden		
Programmeerbare thermostaat			€ 48.05	€ 58.00	644 kWu	10 maanden		
Koken op gas ipv elektrisch			€ 190.00	€ 75.00	210 kWu	31 maanden		
Radiatorfolie aanbrengen			€ 35.00	€ 45.00	500 kWu	9 maanden		

Naam	Moeite eenmalig	Moeite continu	Kosten eenmalig	Besparing euro/jaar	Besparing kWu/jaar	Terugverdien-tijd	Milieu-effecten	Comfort
Koelkast uit bij vakantie			geen	€ 4.20	20 kWu	direct		
Radiatorfolie aanbrengen			€ 35.00	€ 45.00	500 kWu	9 maanden		
Programmeerbare thermostaat			€ 48.05	€ 58.00	644 kWu	10 maanden		
Waterbesparende douchekop			€ 30.00	€ 36.00	400 kWu	10 maanden		
Koken op gas ipv elektrisch			€ 190.00	€ 75.00	210 kWu	31 maanden		

Figure 7: Sorting the recommended choice options on a specific attribute

Furthermore, total savings can be displayed in either Euros or KWh (see Figure 8).



Stap 3: jouw besparingen

Hiernaast zie je de maatregelen die je gekozen hebt!

Geef het totaal weer in ◉ euro ○ kWu

Dit wil ik gaan doen:	
PC energiebeheer inschakelen	€ 67.20
Shirts kort in de droger ipv strijken	€ 12.75
PC uitzetten bij afwezigheid	€ 96.28
Ik kan hiermee per jaar besparen:	**€ 176.23**

Stap 3: jouw besparingen

Hiernaast zie je de maatregelen die je gekozen hebt!

Geef het totaal weer in ○ euro ◉ kWu

Dit wil ik gaan doen:	
PC energiebeheer inschakelen	320 kWu
Shirts kort in de droger ipv strijken	56 kWu
PC uitzetten bij afwezigheid	458 kWu
Ik kan hiermee per jaar besparen:	**834 kWu**

Figure 8: Displaying the total savings in Euros or KWh

Adaptations and explanations

The system is able to automatically change between the variants of each of the features described above based on an internal user model. The user model has two user characteristics (domain knowledge and commitment) which are updated based on process data. Each adaptation has a threshold that triggers the change from 'variant A' to 'variant B'.

Optionally, the system provides an explanation of the adaptive behavior, complete with a reason of why the adaptation was made. The explanation can either be 'generic', with an icon of a light bulb and neutral language, or 'agent-based', with a human-like character and personal language.

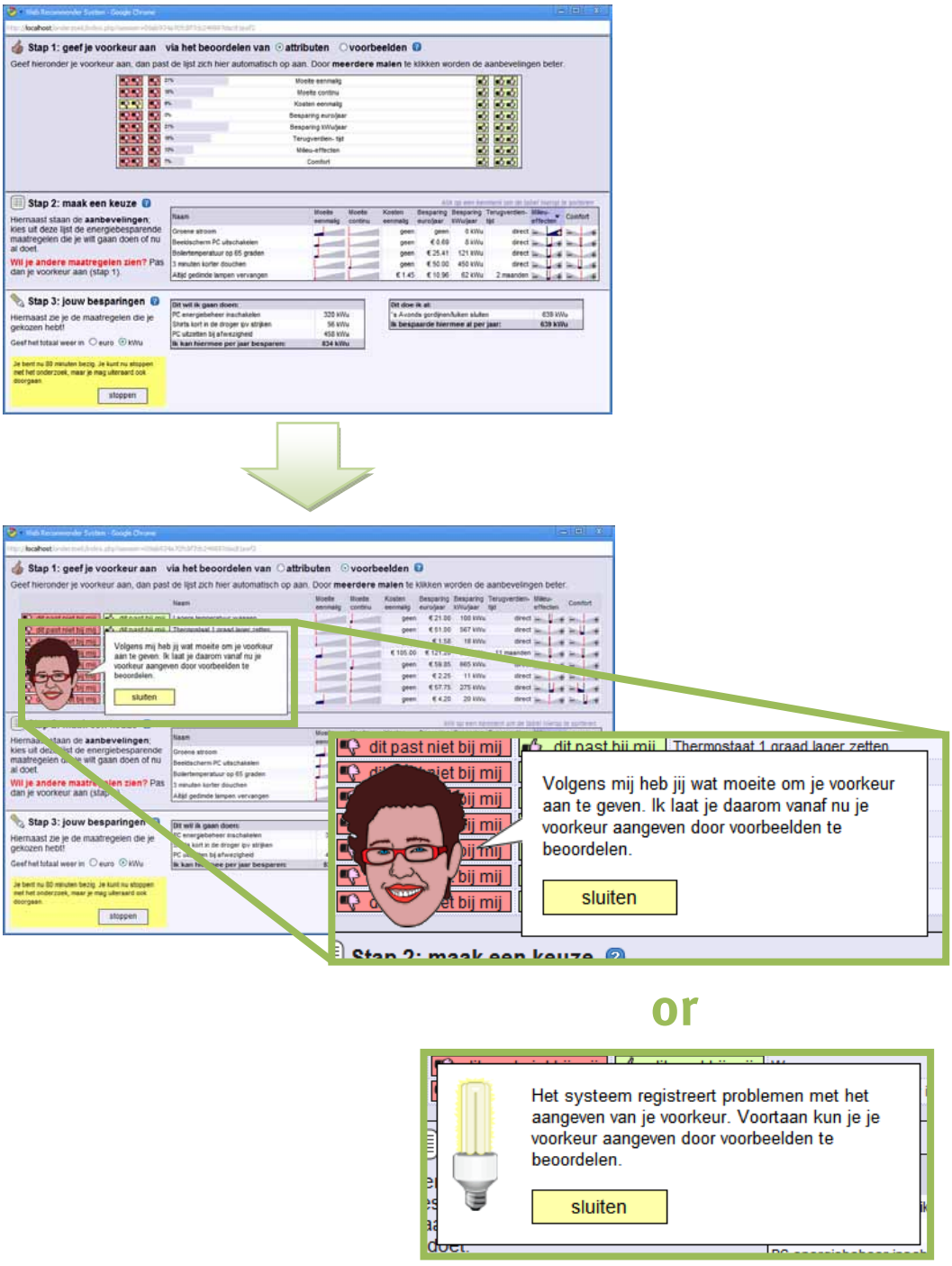


Figure 9: Adaptive behavior; the system automatically changes a feature in the interface based on a user model; changes are explained by an agent (top inset) or a general text-frame (bottom inset)

Making the system adaptive

A detailed description of the adaptive behavior

Steps towards adaptiveness

The adaptive behavior of the Web Recommender System is designed to operate as follows (see Figure 10): The system monitors process data, and uses a set of 'process-rules' to update the values of a user model that has two 'user characteristics': domain knowledge and commitment. Whenever the user model passes a certain threshold, a change is made in the interface[7].

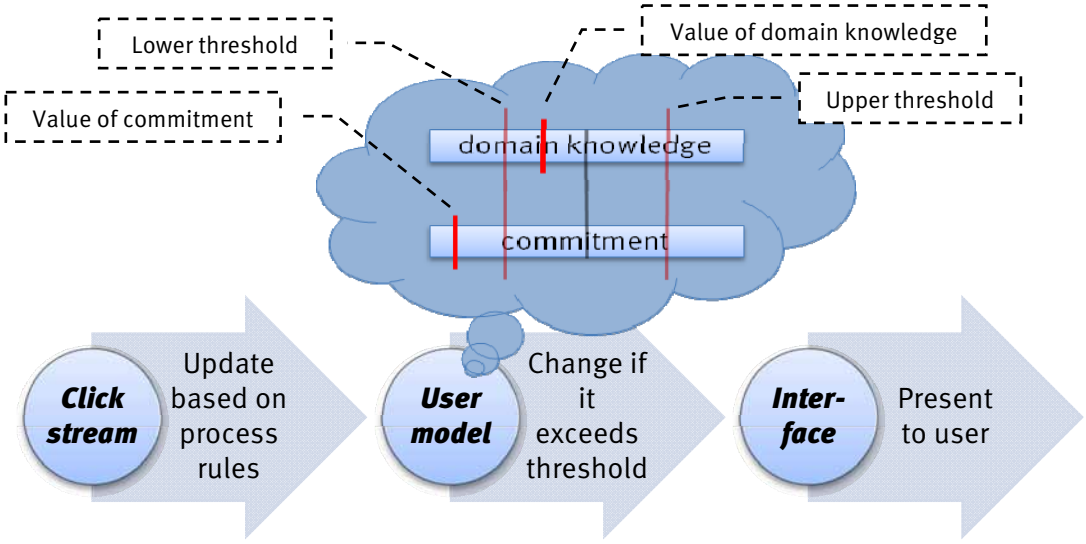


Figure 10: A schematic representation of the adaptive behavior of the Web Recommender System

In order to make adaptiveness work, then, the following things need to be determined:

- The workings of the user model itself.
- The process rules that update the user-model values for domain knowledge and commitment, and the amount with which they update these values.
- The adaptations that can be made by the system, and the thresholds above or below which the adaptations take place.

User model

The user model itself continuously updates the values of the two 'user characteristics': domain knowledge and commitment. The values start at 0 and fluctuate between -1 and +1. When an update applies, the update value is multiplied with the difference between the current value and +1 (for an increase) or -1 (for a decrease). The model is thus biased towards falsification: updates that run against the current beliefs have a higher impact than updates that confirm

[7] Naturally, no change is made if the system is already in the state that the adaptation wants it to change to.

them. Furthermore, the closer the value gets to +1 or -1, the harder it gets to increase or decrease the value respectively. This is illustrated in Figure 11.

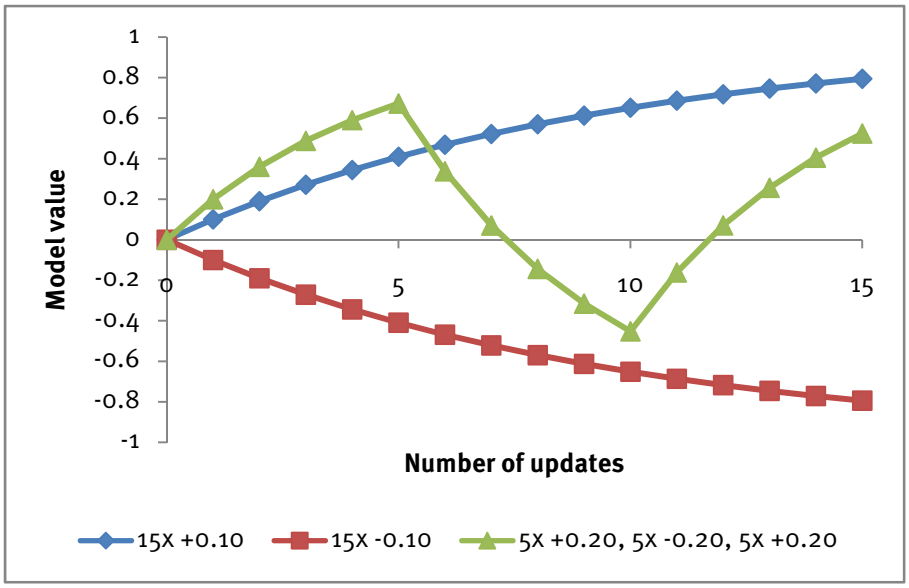


Figure 11: User model value after a certain number of updates

Process-rules

The process-rules indirectly measure the value of the user characteristics. They are based on typical behavior for certain user types. For example, if novices are more likely to decrease their preference weights while experts are more likely to increase them, then a click increasing a preference weight increases the user model value of domain knowledge, while a click decreasing a preference weight decreases the user model value.

Our first experiment links process data to domain knowledge and commitment (measured through questionnaires). Therefore, the results of this experiment were used to determine the optimal values for the process-rules. The procedure of determining these optimal values is described in Appendix G.

Possible adaptations

In the theory chapter of this thesis, we outlined several adaptations to domain knowledge and commitment that could potentially be beneficial.

Beneficial adaptations to domain knowledge

One beneficial adaptation to domain knowledge would be to change the preference elicitation method for different levels of domain knowledge. Experts should get the attribute-based preference-elicitation method, while novices should get the case-based preference-elicitation method. Our first experiment explicitly tests the potential benefit of this adaptation. Another possible adaptation is to present novices with general information about the energy-saving

measures, and to present experts with more detailed, technical information (Alba & Hutchinson, 1987).

Beneficial adaptations to commitment

One beneficial adaptation to commitment would be to sort and highlight different attributes in the interface for different levels of commitment. As people with a low ecological commitment are most interested in personal benefits, the 'comfort' attribute would be their most prominent attribute. People with a high ecological commitment are most interested in environmental benefits, so the 'environmental effects' attribute seems to be the most prominent one for them. Another possible adaptation is to display the savings of chosen measures in Euros (for less committed individuals) or kilowatt-hour (for more committed individuals).

Threshold values

The possible adaptations as discussed above need a certain threshold value above or below which they apply. These thresholds should be chosen carefully, as the interface should not 'flip' too many times, but at the same time should not wait too long adapting the interface when necessary. Our first experiment provided a rich dataset with which we could simulate the adaptive behavior, and find the optimal threshold levels. The procedure of determining these values is described in Appendix G.

Experiments

Designing an experimental plan to test the central thesis argument

In order to test the central thesis argument, two experiments were designed. The first experiment tests the possible benefits of our most consequential adaptation, the preference elicitation method. By randomly assigning participants to a (static) preference elicitation method and measuring their domain knowledge, and, after the interaction, their subjective evaluation of the system, we can check whether matching the preference elicitation method to domain knowledge has the predicted positive effect. Furthermore, this experiment checks whether it is possible to predict both domain knowledge and commitment from process data.

In the second experiment, we use the process data predictors found in the first experiment to make an adaptive version of the system. The experiment tests the effect of adaptiveness, the effect of providing explanations for the adaptations, and the effect of using an agent to present these explanations.

The remaining chapters of this thesis describe the setup and results of the two experiments. Finally, the conclusion chapter reflects on the central thesis argument based on evidence obtained in the experiments.

1st experiment

This chapter addresses the details of the first experiment. This experiment is a precursor for the second experiment, as it tries to prove the feasibility of adapting to ecological knowledge and commitment, and define the details of this adaptive behavior.

The chapter starts with a number of hypotheses that follow from the 'central thesis argument' in the theory chapter. It then gives a detailed definition of the constructed experiment and the measurement tools that are used to test these hypotheses.

Goal of the experiment

Preparation for adaptation

When developing a system that adapts to certain personal characteristics, it may be necessary to first study the link between these personal characteristics and the proposed adaptations (Höök, 1998). Furthermore, such a study could be used to link the personal characteristics to process data. This is exactly the goal of the first experiment of this thesis: providing justification and input for the adaptive system.

Justification for adaptiveness can be provided by testing whether our expectations about the hypothesized differences between novices and experts and between people with different levels of commitment really exist. Specifically, by presenting participants with different preference elicitation methods and measuring their domain knowledge and, after the interaction, their satisfaction, we are able to check whether novices really like the case-based preference elicitation method better than the attribute-based preference elicitation, and vice versa for experts. Moreover, we can check whether people with low and high commitment actually do pay attention to different attributes.

Input for adaptiveness can be provided by first measuring domain knowledge and commitment using questionnaires, and then matching process data to these measures. Significant correlations between process data and these characteristics (e.g. increasing a preference weight is positively correlated to domain knowledge) can be used as input for the definition of rules for a user model that can measure domain knowledge and commitment on the fly (e.g. the user model value of domain knowledge increases slightly whenever a user increases a preference weight). Measuring domain knowledge and commitment with process data increases the value of the adaptive system, as pre-experimental questionnaires will then no longer be required to measure domain knowledge and commitment.

In order to prepare for adaptiveness, this experiment thus needs to measure domain knowledge and commitment with questionnaires, collect process data, manipulate the presented preference elicitation method, and measure post-interaction satisfaction, understandability and perceived usefulness. Based on these requirements, a series of hypotheses is formulated below, and subsequently the setup of an experiment that tests these hypotheses is described.

Hypotheses

Adaptation to domain knowledge is feasible and potentially useful

First and foremost, the successful adaptation of the preference elicitation method to domain knowledge requires that novices and experts actually differ in which method they prefer. Adaptation is only *useful* when novices and experts differ in their preferred interface. Consequently, we can hypothesize:

H1. Novices have a higher satisfaction and perceive the system as more useful when they use the case-based PE method (compared to the attribute-based PE method), while experts have a higher satisfaction and perceive the system as more useful when they use the attribute-based PE method (compared to the case-based PE method).

Finally, in order to adapt to domain knowledge unobtrusively and on the fly, it is required to measure these characteristics during the interaction based on process data. Adaptation is only *feasible* when there exists a correlation between these characteristics and (preferably, several) process data measures. Although we do not endeavor to provide specific hypotheses concerning differences in clicking behavior between novices and experts, we can hypothesize the following:

H2. Novices and experts differ in their clicking behavior, and it is therefore possible to relate the level of domain knowledge to differences in process data.

Adaptation to commitment is feasible and potentially useful

Again, first and foremost, for adaptation to ecological commitment to be *useful*, there needs to be a difference in the choice goals of people with low commitment and high commitment. Specifically, we predict that people with low commitment pay more attention to the personal benefits of energy-saving, while people with high commitment pay more attention to the ecological benefits. This leads to the following hypothesis:

H3. Individuals with low commitment primarily look at personal benefits, i.e. comfort and savings in Euros, while individuals with high commitment primarily look at ecological benefits, i.e. environmental effects and savings in kilowatt-hours.

Furthermore, unobtrusively adapting to ecological commitment is only *feasible* if it is possible to measure the concept during the interaction based on process data. Again, we do not endeavor to provide specific hypotheses concerning differences in clicking behavior, but we can hypothesize the following:

H4. Individuals with different levels of commitment differ in their clicking behavior, and it is therefore possible relate the level of ecological commitment to differences in process data.

Additional predictions

Although of less importance to the main theory of this thesis, two additional predictions can be made based on a review of the existing literature.

First of all, there may be a main effect of user domain knowledge on satisfaction with recommender systems in general. In an experiment evaluating a natural-language based recommender system, Chai et al.(2002) found that novice participants rated the recommender system a lot higher on perceived ease-of-use than a standard menu-based system (no significance values were supplied), while for experts, there was no difference in ease-of use.

Likewise, in an experiment with an online recommender system, Spiekerman (2001) shows that experts were less likely than novices to use the preference elicitation part of the system and more likely to engage in manual search (see Xiao & Benbasat, 2007, p. 170) for a more in-depth analysis of the effect of domain knowledge on the subjective evaluation of recommender systems). This may be due to the fact that "people are most likely to have well-articulated preferences when they are familiar and experienced with the preference object" (Bettman et al., 1998, p. 188). In other words, since experts already know their preference and how it translates in choice options, they have less need for a recommender system's help. This leads us to expect that:

H5. In general, people with a high level of domain knowledge rate the perceived usefulness of the system lower than people with a low level of domain knowledge.

As a cautionary remark, we indicate that these findings seem to contradict other research on online shopping that indicates that web shops are usually more suitable for the expert end of the customer spectrum. Furthermore, the possible match of preference elicitation and domain knowledge may nullify this main effect, as the tailored approach may optimize benefit for both user types.

The second additional prediction concerns a main effect of commitment on interaction with the recommender system. In this respect, Spiekerman (2001) shows that consumers with a

higher involvement in their choice make more extensive use of her recommender system. The increased use of the recommender system as found by Spiekerman may indicate a higher satisfaction and perceived usefulness. Therefore, we predict:

H6. In general, people with a higher ecological commitment will be more satisfied with the system.

And:

H7. In general, people with a higher ecological commitment perceive the system as more useful.

Procedure

Users, system and task

Users were recruited online via Internet forums and featured weblog posts, and digital word-of-mouth[8]. Care was taken that the recruiting websites were both energy-related and general interest. A total of 145 participants started the experiment and got through the pre-experimental questionnaires. 93 of them fully completed the interaction. 89 of these also finished the post-experimental questionnaires[9]. All users were asked to participate using a 'neutral' explanation (appealing to both the environmental and the personal benefits), a request to participate that would "help to make further improvements to the system", and a promise of a small financial reward.

After following the link in the message, participants were informed about the time investment and optional rewards and assigned an ID that could be used at any time to resume the experiment in case of network problems. Subsequently, participants were given several pre-experimental questionnaires measuring demographics, domain knowledge, and commitment, and a step-by-step explanation of the system. Participants were instructed that the goal of the system was to "find new saving measures that match your preference and at the same time catalogue saving measures that you are currently doing already."

Participants were then routed to the actual experiment which they were required to use for at least 10 minutes. After that, participants were allowed to stop the experiment and start on the post-experimental questionnaires measuring satisfaction, perceived usefulness, and

[8] The experiment was posted on the front page of the Olino.org, and DSE.nl weblogs, and on the forums of viva.nl, peakoil.nl and zoom.nl. The experiment was also syndicated over Google Alerts and nujij.nl, and via email by contacting the researchers' personal network (making sure informed contacts were avoided).

[9] Our best guess for the aborted sessions is a lack of interest or time. We found no significant predictors of people prematurely ending the experiment.

understandability. Finally, participants could make a choice of payment method: no payment, participation in a lottery, or a payment of five Euros after registration for an online research panel.

System manipulation

All participants used the Web Recommender System as described in the chapter 'An adaptive recommender system' on page 30 (albeit without the adaptive behavior), and were randomly assigned to one of two conditions: one with a fixed attribute-based preference elicitation method and one with a fixed case-based preference elicitation method.
Participants in this experiment were not able to change the preference elicitation method. Also, the attribute-based preference elicitation did not have the 'increase/decrease more' buttons (buttons with two thumbs), as these buttons were added after experiment 1. The same holds for the 'live help' feature (blue circles with question marks).

Measures

Demographics

For the generality of the results of this experiment, a wide distribution of demographics is preferred. An inspection of the demographics showed that the sample was biased towards males (34 female, 111 male), but had wide distribution of ages (M = 35.7, SD = 11.6), education (12 high school, 23 intermediate vocational education, 59 higher vocational education, 52 university) and occupations (27 students, 104 employed, 14 retired).

Domain knowledge and commitment

Before interaction with the system, 31 five-point scale questions were asked about domain knowledge (18) and commitment (13)[10]. These questions were entered in an exploratory factor analysis, using Generalized Least Squares extraction and Varimax rotation[11]. Initial and extracted communalities were > .30. Based on inspection of the scree plot, 2 factors were extracted, together explaining 36% of the total variance (23% and 13% respectively). After rotation, these factors neatly divided the items in one factor with the domain knowledge items and one with the commitment items. The final factor solution has a KMO-statistic of 0.815, which is well above the required 0.60, and Bartlett's Test of Sphericity shows a significant deviation from the identity matrix, which means that factor analysis is an adequate procedure to use.
4 items were deleted because they had factor loadings < .20 on either factor. In the resulting analysis, 4 items had non-trivial loadings on both factors, and one domain knowledge item

[10] A detailed list of all pre- and post-experimental questions can be found in Appendix F.

[11] Analyses with oblique rotations resulted in uncorrelated factors and did not fit significantly better than this orthogonal rotation.

("All ways of saving energy are basically the same") loaded higher on the commitment factor. The rotated factor solution is displayed in Table 1 below.

Table 1: Factor analysis of the domain knowledge and commitment questions

	Domain knowledge	Commitment
I understand difference between measures	0.76	
I know energy consumption of all devices	0.75	
I know more measures than others	0.73	
I can choose the right measures	0.73	
I know which measures are useful	0.70	
I can make trade-offs between measures	0.63	
I can understand pros and cons of measures	0.59	
I always pay attention to my energy usage	0.58	0.36
I search for extra info about measures	0.58	0.48
Term "energy leakage" is familiar to me	0.53	
Term "ecological footprint" is familiar to me	0.47	
When I implement a measure, it's a conscious trade-off	0.46	
I don't understand most measures	-0.34	-0.23
I think there are better measures	-0.33	
I doubt whether I choose the right measures	-0.26	0.26
Term "carbon cycle" is familiar to me	0.24	
Energy savings gets too much attention		-0.73
People worry too much about the environment		-0.68
When savings cost money this is not annoying		0.65
When savings cost effort this is not annoying		0.63
I encourage other to save energy	0.31	0.61
When savings reduce comfort this is not annoying		0.57
I'm saving energy daily	0.44	0.55
Savings more important than money		0.51
Paying more taxes for environment is not annoying		0.48
Not all savings are worth the effort		-0.35
All measures are eventually the same		-0.23
Eigenvalue	**5.55**	**4.10**

Factor-scores were saved per participant using the regression method. After deleting one outlier, both variables are normally distributed based on skewness and kurtosis tests[12], except that commitment is negatively skewed ($z_{skewness}$ = -3.59). This means that there are some negative outliers on the commitment scale; some participants had an uncharacteristically low commitment.

Concluding, the analysis provided two uncorrelated, normally distributed measures of domain knowledge and commitment, based on factor scores as defined by the loadings displayed above.

[12] Kolmogorov-Smirnov and Shapiro-Wilk tests are significant, but these tests are notoriously sensitive.

Satisfaction with the system

After interaction with the system, satisfaction with the system was measured using the five general items of the QUIS[13]. The nine-point scaled items were summed (M = 26.0, SD = 8.06) to a scale which showed a Cronbach's alpha of 0.83 and a normal distribution, according to skewness, kurtosis Kolmogorov-Smirnov and Shapiro-Wilk tests.

Perceived usefulness, understandability, and satisfaction with the chosen measures

In order to cover as many aspects of satisfaction as possible, the post-experimental questionnaires included 21 five-point scale questions covering the subjective impact of the system on saving behavior, ease of use, clarity of and satisfaction with the recommendation aid, and satisfaction with the chosen energy-saving measures.

These questions were entered in an exploratory factor analysis, using Maximum Likelihood extraction[14] and Oblimin rotation (δ = -.5)[15]. One item was deleted due to extreme multicollinearity, another item was deleted due to low initial communality. For the remaining items, all initial communalities were > .30, but extracted communalities of 5 items did not meet this criterion[16].

Based on inspection of the scree plot, 3 factors were extracted that together explained 47% of the variance (31%, 10% and 6% respectively). After rotation, items divided among these factors with 6 items loading on two factors simultaneously. The factors were interpreted to entail the concepts 'perceived usefulness of the system', 'understandability of the interaction' and 'satisfaction with the chosen measures'. The rotated factor solution is displayed in Table 2 below. This solution has a KMO-statistic of 0.816, which is well above the required 0.60, and Bartlett's Test of Sphericity shows a significant deviation from the identity matrix, which means that factor analysis is an adequate procedure to use.

[13] The questionnaire can be found at http://hcibib.org/perlman/question.cgi?form=QUIS. We excluded item 4 (inadequate power – adequate power), because they raised questions during pretesting.

[14] Maximum Likelihood extraction was used because Generalized Least Squares extraction resulted in a non-significant Goodness-of-fit.

[15] Oblique rotation was used because the factors were thought to be conceptually related. The analysis provided significantly correlated factors, and the factor scores in the sample were also significantly correlated.

[16] This is due to the Maximum Likelihood extraction method, which reduces the impact of variables with low initial communalities.

Table 2: Factor analysis of the subjective evaluation questions

	Usefulness	Under-standability	Satisfaction with measures
I would use the system more often	0.73		0.26
The recommendations fitted my preference	0.71		
I make better choices with the system	0.69		
The system was useless	-0.69		
I would recommend the system to others	0.68		
The system understood my preference	0.66	0.21	
The system made bad recommendations	-0.63		
The system made me more energy-conscious	0.53		
The system restricted my choice freedom	-0.23		
The system was easy to use		0.86	
It was easy to state my preference		0.74	
The system confused me		-0.65	-0.23
It was easy to compare measures	0.22	0.57	
I didn't understand the system at all	-0.21	-0.51	
I understood how to indicate my preference		0.49	
The chosen measures fit my preference			0.86
I like the measures I've chosen		0.26	0.56
I think I chose the best measures			0.39
How many measures will you implement			0.36
Eigenvalue	**4.80**	**3.82**	**2.68**
	Inter-factor correlation		
Usefulness		0.324	0.288
Understandability	0.324		0.294
Satisfaction with measures	0.288	0.294	

Factor-scores were saved per participant using the regression method. All three variables are normally distributed based on skewness and kurtosis tests[17], except that both 'usefulness' and 'satisfaction with measures' are negatively skewed ($z_{skewness}$ = -2.57 and $z_{skewness}$ = -3.36 respectively). This means that there are some negative outliers on these scales; some participants had an uncharacteristically low perceived usefulness or satisfaction with the chosen measures.

Concluding, the analysis provided three correlated, normally distributed measures of 'perceived usefulness of the system', 'understandability of the interaction' and 'satisfaction with the chosen measures', based on factor scores as defined by the loadings displayed above.

[17] Kolmogorov-Smirnov and Shapiro-Wilk tests are significant, but these tests are notoriously sensitive.

Results of the 1st experiment

This chapter enunciates the results of the first experiment. As this experiment was designed as a precursor to the second experiment, the results are addressed in a fashion that helps the reader understand its implications for the second experiment.
Specifically, the chapter first shows the feasibility and probable effect of adaptation to the level of domain knowledge. It indicates a set of predictors of domain knowledge, show that experts and novices differ in their preferred way of explicating their internal preferences, and takes note of some other things that experts or novices typically want in the choice environment (e.g. other things that we could automatically change for them in an adaptive system).
Furthermore, the chapter treats adaptation to commitment in a similar way. It indicates a set of predictors of ecological commitment, shows that committed individuals differ in their choice goals compared to less committed individuals, and makes note of some other things that committed or uncommitted individuals typically want in the choice environment.

The effect of a matching preference elicitation method

On satisfaction with system and perceived usefulness

The hypotheses H1, H5, H6 and H7 were tested by performing linear regressions using our manipulation of the preference elicitation method (PE-method; attribute-based versus case-based), domain knowledge and commitment as predictors, and satisfaction with the system, perceived usefulness, understandability, and satisfaction with the chosen measures as dependent variables. Linear regressions were used because the sum scale of satisfaction and the factor scores of the other measures were constructed as an interval scale. Predictors were either nominal (PE-method) or interval scales (domain knowledge and commitment).

Our main hypothesis in this section (H1) predicts that users who experience a PE-method that is matched their domain knowledge have a higher satisfaction and perceived usefulness than users who experience a PE-method that is not matched to their domain knowledge. This means that we are looking for the effect of the interaction between PE-method and domain knowledge. Specifically, if PE-method is coded -1 for case-based PE and +1 for attribute-based PE, then a higher value of (PE-method * domain knowledge) should lead to a higher satisfaction and perceived usefulness, because this value is higher exactly when novices (negative domain knowledge) experience the case-based PE method (negative PE-method) and when experts (positive domain knowledge) experience the attribute-based PE method (positive PE-method).

Furthermore, if there is no main effect of PE-method, this would mean that *neither* of the PE-methods has a higher satisfaction (or perceived usefulness) for *all* users in general, but that the preferred PE-method solely depends on the domain knowledge of the user.

Such an interaction effect without a main effect is called a 'double dissociation'. If and only if a double dissociation is found, we can conclude that it would be best practice to give experts the attribute-based PE method and novices the case-based PE method, as predicted in H1.

Predicting 'satisfaction with the system'

Table 3 presents the results of the regression on satisfaction.

Table 3: Predicting satisfaction (adjusted R^2 = .165)

	Estimate	Std. Error	t	Partial n^2
Intercept	25.364	0.815	31.111***	0.923
PE-method	0.726	0.815	0.890	0.010
Domain knowledge	-1.155	0.864	-1.337	0.022
Commitment	3.236	0.865	3.739***	0.147
Domain knowledge* PE-method	1.985	0.864	2.297*	0.061
Commitment* PE-method	-1.153	0.865	-1.332	0.021

*$p < .05$ **$p < .01$ ***$p < .001$

First of all, H6 predicted that committed individuals are more satisfied with the system than are less committed individuals. Our experiment confirmed this hypothesis with a large-sized significant main effect of commitment on satisfaction.

More importantly, H1 predicted that experts are more satisfied with the attribute-based preference elicitation method while novices are more satisfied with the case-based preference elicitation method. This hypothesis was supported with a medium-sized significant effect of the interaction between domain knowledge and preference elicitation method on satisfaction, and the absence of a main effect of PE-method. The predicted double dissociation is displayed in Figure 12.

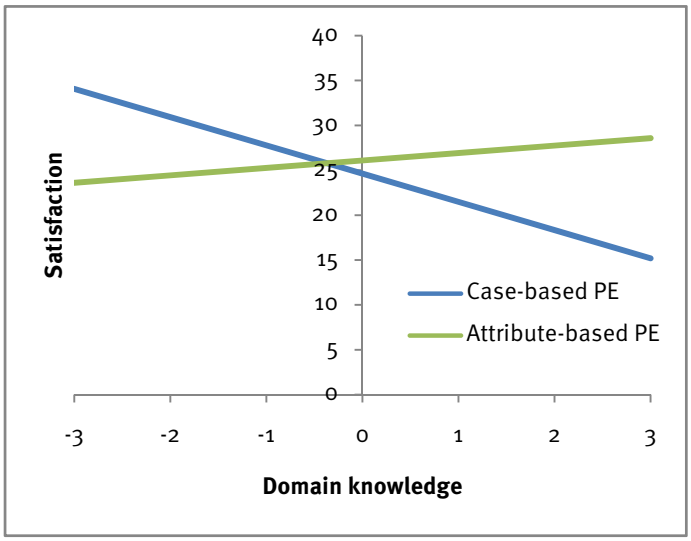


Figure 12: Value of satisfaction with the different PE-methods for participants with a certain level of domain knowledge

Predicting perceived usefulness

Results of the regression on perceived usefulness are presented in Table 4.

Table 4: Predicting perceived usefulness (adjusted R^2 = .265)

	Estimate	Std. Error	t	Partial n^2
Intercept	-0.069	0.092	-0.750	0.007
PE-method	-0.150	0.092	-1.634	0.033
Domain knowledge	-0.054	0.096	-0.557	0.004
Commitment	0.427	0.097	4.416***	0.200
Domain knowledge* PE-method	0.302	0.096	3.133**	0.112
Commitment* PE-method	-0.136	0.097	-1.409	0.025

*$p < .05$ **$p < .01$ ***$p < .001$

H7 predicted that committed individuals perceive the system as more useful than less committed individuals. We confirmed this hypothesis with a large significant main effect of commitment on perceived usefulness.

H1 not only predicted that novices and experts that use the 'right' preference elicitation method are more satisfied with the system, but also that they perceive the system as more useful. This hypothesis was again supported with a medium-sized significant effect of the interaction of domain knowledge and PE-method on perceived usefulness, and the absence of a main effect of PE-method. This means that novices rated the system with case-based preference elicitation as more useful, while experts rated the system with attribute-based preference elicitation as more useful. The double dissociation is displayed in Figure 13.

H5 predicted that experts judge the system as less useful than novices, regardless of the preference elicitation method used. This hypothesis was not confirmed, as there was no significant main effect of domain knowledge on perceived usefulness.

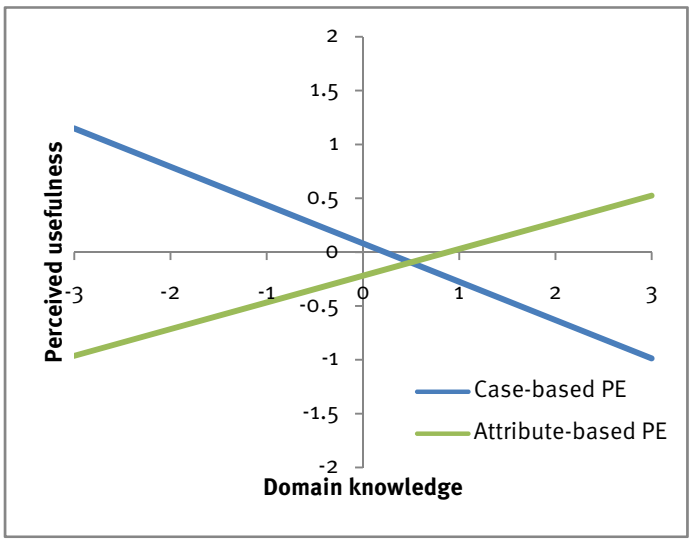


Figure 13: Value of perceived usefulness with the different PE-methods for participants with a certain level of domain knowledge

Additional observations

Predicting understandability, satisfaction with the chosen measures and total amount of energy saved

Predicting understandability

Although understandability is not the subject of any specific hypothesis, it was included in the analysis to see if other usability aspects were influenced by the measured user characteristics or the presented preference-elicitation method. Results of the regression on understandability are presented in Table 5.

Table 5: Predicting understandability (adjusted R^2 = .051)

	Estimate	Std. Error	t	Partial n^2
Intercept	-0.032	0.102	-0.311	0.001
PE-method	0.266	0.102	2.608*	0.080
Domain knowledge	-0.118	0.107	-1.099	0.015
Commitment	0.157	0.108	1.462	0.027
Domain knowledge* PE-method	0.050	0.107	0.471	0.003
Commitment* PE-method	-0.070	0.108	-0.655	0.005

*$p < .05$ **$p < .01$ ***$p < .001$

Predicting understandability, we found a medium-sized significant effect of condition. This means that, on average, the attribute-based preference elicitation method is more understandable than the case-based preference elicitation. This result can be explained by the fact that the interaction with the attribute-based preference elicitation is a straightforward specification of attribute weights and it gives an unambiguous display of the user's preference. The case-based preference elicitation method, on the other hand, 'disguises' the specification of attribute weights in a less understandable critiquing of examples.

Although this effect holds for any level of expertise, it is especially surprising for novices, because although they perceive the case-based preference elicitation as more useful and satisfying, they still find the attribute-based preference elicitation more understandable.

Predicting satisfaction with the chosen measures

Besides satisfaction with the system, we also included a measure of satisfaction with the chosen energy-saving measures in our analysis in order to check whether the measured user types or the presented preference elicitation method would have an effect beyond the interaction with the system itself.

A regression with preference elicitation method, domain knowledge and commitment as predictors provided no significant results. Consequently, we predicted that such an effect could be mediated by the other measures. The results of a regression with 'satisfaction with the system', 'perceived usefulness' and 'understandability' as predictors are shown in Table 6 below.

Table 6: Predicting satisfaction with the chosen measures (adjusted R^2 = .221)

	Estimate	Std. Error	t	Partial n^2
Intercept	-1.201	0.453	-2.653**	0.076
Satisfaction with system	0.046	0.017	2.700**	0.078
Perceived usefulness	0.035	0.124	0.280	0.001
Understandability	0.079	0.116	0.676	0.005

*$p < .05$ **$p < .01$ ***$p < .001$

We found a medium-sized significant effect of 'satisfaction with the system' on 'satisfaction with the chosen measure'. This means that the higher the satisfaction with the system, the

higher the satisfaction with the chosen measures. This is an interesting result which suggests that in general, the satisfaction with a recommender system can reflect on the items chosen/purchased with help of the system.

Predicting total amount of energy saved

Looking at the energy-saving measures chosen by our participants during the interaction is another way to check whether measured user types or preference elicitation method had any influence beyond the interaction with the system. Specifically, we measured the total amount of energy saved (in kilowatt-hours) with the measure that each user had chosen to implement. Table 7 presents the results of the regression with preference elicitation method, commitment and domain knowledge as predictors and the total amount of energy saved as dependent variable.

Table 7: Predicting total amount of energy saved (adjusted R^2 = .058)

	Estimate	Std. Error	t	Partial n^2
Intercept	2293.56	292.886	7.831***	0.000
PE-method	-185.23	292.886	-0.632	0.529
Domain knowledge	-19.24	311.729	-0.062	0.951
Commitment	879.50	311.927	2.820**	0.006
Domain knowledge* PE-method	163.30	311.729	0.524	0.602
Commitment* PE-method	311.90	311.927	1.000	0.320

*$p < .05$ **$p < .01$ ***$p < .001$

We found a medium-sized significant effect of commitment on total amount of energy saved. This means that more committed individuals chose to save more energy.

There was no significant effect of the interaction between domain knowledge and PE-method like we found for satisfaction and perceived usefulness. This means that although matching the PE-method increased user satisfaction and perceived usefulness of the system, it did not lead people to save more energy. This is not necessarily a disappointing result, as not all participants have a goal of saving energy. Furthermore, as a higher satisfaction and perceived usefulness increase the chance of future use of the system (Xiao & Benbasat, 2007), we predict that the long run results of a tailored interface may indicate increased energy savings.

Process data predictors

Of domain knowledge and commitment

In this experiment, we measured domain knowledge and commitment using 31 questions. Taking the time to ask users these questions is a luxury that 'real life' recommender system often cannot afford. Consequently, it would be very practical if we were able to find significant differences in interaction behavior between participants with different levels of domain

knowledge and commitment. Such relations between the user characteristics and process data could then be used to define rules for updating a user model of these characteristics in an adaptive system.

Furthermore, if the predictors of commitment show that people with low commitment primarily prefer (measures with) the attributes comfort and savings in Euros, while people with high commitment primarily prefer (measures with) environmental effects and savings in kilowatt-hours, this confirms H3.

Click frequencies

The most straightforward process data measure is the frequency with which every type of click is performed by the user. We discerned the following click types: increasing an attribute weight, decreasing an attribute weight, selecting an item to see its details, choosing an item or indicating an item as 'already applied', sorting the recommendations, changing the information type (general or details) and changing the display unit of total savings (Euros or KWh). H2 and H4 predict that the frequencies of these click types can be used to predict domain knowledge and commitment respectively.

We performed stepwise regressions with the frequencies of the click types as predictors to determine if differences in certain click type frequencies could predict domain knowledge (Table 8) and commitment (Table 9).

Table 8: Predicting domain knowledge based on click frequencies (adjusted R^2 = .179)

	Estimate	Std. Error	t	Partial n^2
Intercept	-0.526	0.168	-3.135**	0.095
Increase preference weight	0.003	0.001	2.683**	0.071
Choose item[18]	-0.041	0.018	-2.207*	0.049
Indicate 'already doing this'	0.037	0.010	3.901***	0.139

*$p < .05$ **$p < .01$ ***$p < .001$

We found that participants with a higher level of domain knowledge more frequently increase attribute weight, choose less measures, and mark more measures as "already applied".

Table 9: Predicting commitment based on click frequencies (adjusted R^2 = .051)

	Estimate	Std. Error	t	Partial n^2
Intercept	-0.080	0.122	-0.656	0.004
Choose item	0.038	0.015	2.486*	0.061

*$p < .05$ **$p < .01$ ***$p < .001$

We found that more-committed individuals choose more measures than less-committed individuals.

[18] As users were allowed to change their minds on their choices, we corrected the variables 'choose item' and 'indicate already doing this' for the items that were 'unchosen' later on.

Chosen measures

In the experiment, users were asked to indicate which energy-saving measures they already applied, and which measures they were considering to implement. The items specified as such provide another set of possible predictors for domain knowledge and commitment, especially the minimum, maximum, mean and sum of the attribute values of these chosen energy-saving measures. These predictors were entered into stepwise regressions with domain knowledge (Table 10) and commitment (Table 11) as dependent variables.

Table 10: Predicting domain knowledge based on chosen measures (adjusted R^2 = .229)

	Estimate	Std. Error	t	Partial n^2
Intercept	-0.176	0.204	-0.866	0.008
Sum of savings (kWh) of measures already applied	0.000	0.000	5.193***	0.223
Mean continuous effort of measures already applied	-0.120	0.043	-2.797**	0.077
Sum of comfort of measures already applied	-0.006	0.003	-2.184*	0.048

*$p < .05$ **$p < .01$ ***$p < .001$

We found that people with a higher level of domain knowledge apply measures with higher sum of KWh savings, a lower average level of continuous effort and a lower sum of comfort.

Table 11: Predicting commitment based on chosen measures (adjusted R^2 = .307)

	Estimate	Std. Error	t	Partial n^2
Intercept	-0.526	0.267	-1.970	0.041
Maximum of savings (kWh) of chosen measures	0.000	0.000	3.772***	0.135
Maximum of environmental effects of chosen measures	0.029	0.010	2.894**	0.084
Sum of environmental effects of measures already applied	0.013	0.003	4.161***	0.160
Minimum of cost once of measures already applied	-0.010	0.004	-2.632**	0.071
Sum of cost once of measures already applied	0.000	0.000	2.621**	0.070
Maximum of comfort of measures already applied	-0.035	0.013	-2.734**	0.076

*$p < .05$ **$p < .01$ ***$p < .001$

We found that participants with a higher level of commitment choose measures with higher maximum of KWh savings and environmental effects. They also already apply measures with a higher sum of environmental effects, a lower minimum of one-time costs, a higher sum of one-time costs and a lower maximum comfort.

Average attribute weights

During the interaction, the system continuously records and updates the users' weights for each attribute. Whenever the user makes a choice (either indicating an intention to apply this measure or the fact that it is already being applied), one may derive from this choice that the attribute weights are in some way 'correct', or at least good enough to provide measures that are of interest to that user.

By taking the average of these 'correct' attribute weights we can construct an 'average user preference'. These average weights can be used to predict domain knowledge and commitment, and are therefore entered in stepwise regressions.

The final regression for domain knowledge provided no significant results. The results of the final regression for commitment are displayed in Table 12 below.

Table 12: Predicting commitment based on average attribute weights (adjusted R^2 = .113)

	Estimate	Std. Error	t	Partial n^2
Intercept	-0.106	0.222	-0.476	0.002
Avg. preference for low continuous effort	-2.518	1.255	-2.007*	0.041
Avg. preference for positive environmental effects	3.438	1.033	3.327***	0.104

*$p < .05$ **$p < .01$ ***$p < .001$

We found that more committed individuals have a lower average preference for low continuous effort and a higher average preference for environmental effects. In general, people with a higher commitment find low continuous effort less important and positive environmental effect more important.

Demographics

We also tested whether demographics could predict domain knowledge and commitment. Gender, age, occupation and education were entered in stepwise regressions. The results of the final regression for domain knowledge are displayed in Table 13.

Table 13: Predicting domain knowledge based on demographics (adjusted R^2 = .194)

	Estimate	Std. Error	t	Partial n^2
Intercept	-0.625	0.159	-3.941***	0.099
Gender	0.957	0.166	5.750***	0.190
Occupation[19]				
Retired	0.151	0.239	0.632	0.003
Student	-0.655	0.183	-3.588***	0.084

*$p < .05$ **$p < .01$ ***$p < .001$

[19] Using 'employed' as the reference category.

We found that men have on average more domain knowledge about energy-saving than women. Furthermore, compared to employed people, students have less domain knowledge about energy-saving measures. The final regression for commitment provided no significant results.

Conclusions about using process data

Based on evidence in the paragraphs above, we can conclude that significant relations between domain knowledge and commitment on the one hand, and process data on the other hand, exist. Although it is far from possible to predict these user characteristics without any error, this confirms hypotheses H2 and H4: It is to some extent possible to measure domain knowledge and commitment on the fly.

Furthermore, the chosen measures and preference weights predicting commitment indicate that committed individuals like (measures with) high environmental effects and KWh savings, while less-committed individuals like (measures with) a higher level of comfort. This confirms H3.

Conclusion

Adapting to domain knowledge and commitment is promising

All hypotheses of experiment 1 were confirmed, except for H5, which predicted that people with a high level of domain knowledge rate the perceived usefulness of the system lower than people with a low level of domain knowledge. This prediction was derived from findings by Chai et al. (2002) and Spiekermann (2001), and the added value of the recommender systems in these studies came to the expense of a more elaborate dialogue with the system. The added value in our experiment is apparent even for experts, as it is to our knowledge the only comprehensive overview of energy-saving measures available on the Internet.

Participants that used a preference elicitation method that was matched to their level of domain knowledge were more satisfied with the system and judged it to be more useful than participants that used the preference elicitation method that did not match their level of domain knowledge. Furthermore, we found several process data predictors that can be used to measure users' level of domain knowledge on the fly.

We also confirmed that participants with a higher commitment focused on environmental aspects of energy-saving (positive environmental effects and higher KWh savings) while participants with a lower commitment focused on personal aspects (more comfort and less effort). For commitment, too, we found several process data predictors.

Concluding, the basic requirements for making an adaptive version of the system are met, and such a system has potential merit over the static version.

Discussion

In the theory chapter we argued that, besides the preference elicitation method, the amount of information detail would also be a beneficial adaptation to domain knowledge. In the current experiment, users were allowed to switch between general and detailed information (used by 61 of 98 participants), but the manually chosen information type was not significantly correlated with the user's level of domain knowledge. However, based on the existing literature on user expertise, we still predict that adapting the amount of information detail to the users' level of domain knowledge may be beneficial.

Furthermore, we argued in the theory chapter that people with different levels of commitment would prefer the recommended measures to be sorted on different attributes (comfort for low commitment, environmental effects for high commitment), and would prefer a different display of the total savings (Euros for low commitment, kilowatt-hours for high commitment). In the current system it was possible to sort on any attribute (used by 39 of 98 participants), as well as to switch the display of totals between Euros and kilowatt-hours (used by 40 of 98 participants), but neither of these actions was significantly correlated with the user's level of commitment. However, based on the reasoning that people with low commitment want to save money while people with high commitment want to save energy, we believe that these adaptations can be a beneficial.

As a final remark, we contend that the relation between our user characteristics and process data is far from perfect. This means that although it is possible to measure the characteristics on the fly, these measures will have substantial error. Such error may reduce the beneficial effect of an adaptive system.

2nd experiment

This chapter addresses the details of the second experiment. This experiment implements the adaptiveness proposed as a result of the first experiment and tests whether the adaptiveness has the predicted beneficial effect. It furthermore tests the effects of a human-like agent that explains the adaptive behavior of the system.
The chapter starts with a number of hypotheses that follow from the 'central thesis argument' in the theory chapter. It then gives a detailed definition of the constructed experiment and the measurement tools that are used to test these hypotheses.

Goal of the experiment

Testing adaptiveness, explanations and agents

The results of the first experiment support the argument that people with different levels of domain knowledge and commitment should be given different interfaces, and that these user characteristics can to some extent be measured on the fly using process data.

This suggests that adapting to domain knowledge and commitment using process data is, in theory, both feasible and beneficial. However, it remains unknown whether the measurement of user characteristics through process data is sufficiently accurate and precise, and whether users accept the on-the-fly adaptation. Imprecise or inaccurate measurements and unpredictable transformations of certain interface elements could potentially reduce the understandability, satisfaction, and perceived usefulness of the system. In order to find out whether an adaptive system is really beneficial, one would need to test an actual adaptive system with real users. The second experiment therefore tests a system that adapts to the user as proposed in experiment 1.

As explained in the theory chapter of this thesis, the adaptive behavior of a system might confuse users, thereby nullifying any beneficial effects of adaptation. Explaining the adaptive behavior may prevent this confusion, thereby increasing satisfaction. Furthermore, we predict that using a human-like agent to explain the adaptations may be the most beneficial way to explain the adaptive behavior.

In order to test the adaptiveness, explanations and agents, we present each participant with one of four different systems: a 'baseline' system without adaptive behavior, an adaptive system that does not provide explanations, an adaptive system that provides explanations, and an adaptive system that provides explanations using a human-like agent.

Based on predicted differences in the subjective evaluation of these four systems, a series of hypotheses is formulated below, and subsequently the setup of an experiment that tests these hypotheses is described.

Hypotheses

Adaptation with explanation works

First and foremost, we expect adaptiveness to have beneficial effects on the interaction with the system. Specifically, we can hypothesize:

H8. Participants using the adaptive system with explanations have a higher satisfaction and perceive the system as more useful than participants using the system without adaptiveness.

And:

H9. Participants using the adaptive system with agent-based explanations have a higher satisfaction and perceive the system as more useful than participants using the system without adaptiveness.

Adaptation is confusing without explanation

We predict that the adaptive system without explanation confuses participants, which reduces understandability, satisfaction and perceived usefulness. We therefore hypothesize:

H10. Participants using the adaptive system without explanations judge the system in general and the adaptation itself to be less understandable, less satisfying and less useful than participants using the system without adaptiveness.

Adaptation works best with a human-like agent

Finally, we predict that a human-like agent is most suited to explain the adaptive behavior of the system. Specifically, we hypothesize:

H11. Participants using the adaptive system with agent-based explanations contend that the system provides more personal help and are more willing to accept the adaptive behavior compared to participants using the other systems.

Procedure

Users, system and task

Users were recruited online via Internet forums and featured weblog posts, and digital word-of-mouth[20]. Care was taken that the recruiting websites were both energy-related and general interest. A total of 229 participants started the experiment and got through the pre-experimental questionnaires. 149 of them fully completed the interaction. 131 of these also finished the post-experimental questionnaires[21]. All users were asked to participate using a 'neutral' explanation (appealing to both the environmental and the personal benefits), a request to participate that would "help to make further improvements to the system", and a promise of a small financial reward.

[20] The experiment was posted on the front page of the Olino.org, and DSE.nl weblogs, and on the forums of fok.nl, blijfpositief.nl and wuz.nl (part of telegraaf.nl). The experiment was also syndicated via email by contacting the researchers' personal network (making sure informed contacts were avoided).

[21] Again we found no significant predictors of people prematurely ending the experiment.

The procedure of participation was similar to experiment 1, with some notable exceptions. First of all, domain knowledge and commitment questionnaires consisted of only 4 questions each. The answers to these questions were not used to construct the user models of these characteristics, but we used these questions to get a rough estimate of the accuracy and precision of the constructed user models.

Furthermore, the step-by-step explanation of the system and the goal of the experiment were shown during the interaction with the system, not preceding it. Participants were required to use the system for at least 5 instead of 10 minutes[22]. The post-experimental questionnaires included extra items to measure the effect of adaptiveness. Finally, we removed the lottery option from the payment methods.

System manipulation

All participants used the Web Recommender System as described in the section 'Description of the system' on page 31. Three changes were made in the interface compared to experiment 1: We implemented on-demand help buttons, 'increase/decrease more' buttons in the attribute based preference elicitation (the double thumbs; see Figure 4 on page 33), and the option to switch preference elicitation method manually.

The option to switch the preference elicitation method was introduced to make sure that there was no difference in the interactive capabilities of the static and the adaptive systems. The on-demand help buttons and the 'increase/decrease more' buttons were included based on feedback gathered in experiment 1.

Participants were randomly assigned to one of four 'system types'. The 'static' system is our baseline condition without adaptiveness, in which the user controls the entire interaction. The 'adaptive, no explanations' system makes adaptations according to our user model[23], but does not explain the adaptations. In the 'explain' condition, our system adapts to the user and also explains the adaptations and the reason for performing them in a neutral tone. In the 'adaptive with agent-based explanations' condition, a similar explanation is given by a human-like agent that uses a personal tone.

The initial state of the adaptive aspects was also randomized in a 2x2 fashion: high domain knowledge (attribute-based preference elicitation and detailed information) versus low domain knowledge (case-based preference elicitation and overview information), and high commitment (total savings in KWh) versus low commitment[24] (total savings in Euros). Note that all adaptive features could also manually be changed by the users themselves, in all of the conditions.

[22] We reduced the minimal interaction time because several participants in the first experiment complained that they were 'done' before the time period of 10 minutes ended.

[23] Details on the actual implementation of the adaptive system and its user model based on data from experiment 1 can be found in Appendix G.

[24] Note that the sorting was initially on the 'name' attribute in all conditions.

Measures

Demographics

For the generality of the results of this experiment, a wide distribution of demographics is preferred. An inspection of the demographics showed that the sample was biased towards males (36 female, 96 male), but had wide distribution of ages (M = 40.1, SD = 12.0), education (12 high school, 22 intermediate vocational education, 62 higher vocational education, 36 university) and occupations (18 students, 95 employed, 19 retired).

Domain knowledge and commitment

Although the adaptive system measures domain knowledge and commitment using click stream data, 4 high-loading items of each of the questionnaires about domain knowledge and commitment from experiment 1 were also included in experiment 2 to get a basic understanding of the accuracy of the user models of the adaptive system. The Chronbach's α reliability of a scale of the four commitment items was only .217. A factor analysis did not provide a robust solution; the KMO-statistic was 0.549, which is below the acceptable 0.60 level. This means that the items had too much 'uniqueness' to summarize them in a single construct; apparently, using only four items is insufficient to measure commitment.

The domain knowledge items formed a coherent factor explaining 52.4% of the variance of the four items (see Table 14). This solution has a KMO-statistic of 0.795, which is well above the required 0.60, and Bartlett's Test of Sphericity shows a significant deviation from the identity matrix, which means that factor analysis is an adequate procedure to use.

Table 14: Factor analysis of domain knowledge questions

	Domain knowledge
I understand difference between measures	0.72
I know energy consumption of all devices	0.59
I know more measures than others	0.79
I know which measures are useful	0.77
Eigenvalue	**5.55**

Factor-scores were saved per participant using the regression method. The variable was normally distributed based on skewness and kurtosis tests.

Satisfaction with the system

After interaction with the system, satisfaction with the system was measured using the five general items of the QUIS. The nine-point scaled items were summed (M = 25.7, SD = 7.35) to a scale which had a Cronbach's alpha of 0.82 and a normal distribution, according to skewness, kurtosis Kolmogorov-Smirnov and Shapiro-Wilk tests.

Perceived usefulness, understandability, and satisfaction with the chosen measures

Factor analysis of the 'perceived usefulness', 'understandability' and 'satisfaction with the chosen measures' provided results similar to experiment 1, with a KMO-statistic of 0.867 and a significant Bartlett's Test. Factor scores were saved using the regression method.

Perceived personal help

After interaction with the system, 8 five-point scale questions were asked about the extent to which participants perceived the system to provide personal help. These questions were entered in an exploratory factor analysis, using Maximum Likelihood extraction. Initial and extracted communalities were > .50. Based on inspection of the scree plot, 1 factor was extracted, explaining 59% of the total variance. One item was deleted because of a low communality. The factor solution is displayed in Table 15 below. This solution has a KMO-statistic of 0.915, which is well above the required 0.60, and Bartlett's Test of Sphericity shows a significant deviation from the identity matrix, which means that factor analysis is an adequate procedure to use.

Table 15: Factor analysis of the personal help questions

	Perceived personal help
The system thinks my way	0.81
The system is helpful	0.79
The system is smart	0.78
The system adapts to me	0.77
The system does what I want	0.76
The system and I were a team	0.76
The system gave personal help	0.73
Eigenvalue	**4.16**

Factor-scores were saved using the regression method. The measure is normally distributed based on Kolmogorov-Smirnov and Shapiro-Wilk tests.

Concluding, the analysis provided a normally distributed measure of 'personal help', based on factor scores as defined by the loadings displayed above.

Acceptance and understanding of adaptive behavior

After the interaction, 7 five-point scale questions were asked about the acceptability and understandability of the adaptiveness of the system. These questions were only asked to participants that actually experienced an adaptation. The questions were entered in an exploratory factor analysis, using Maximum Likelihood extraction. Initial and extracted communalities were > .40. Based on inspection of the scree plot, one factor was extracted, with 55% of the total variance. Two items were deleted because of low communalities. The

factor solution is displayed in Table 16 below. This solution has a KMO-statistic of 0.825, which is well above the required 0.60, and Bartlett's Test of Sphericity shows a significant deviation from the identity matrix, which means that factor analysis is an adequate procedure to use.

Table 16: Factor analysis of the adaptive behavior questions

	Acceptance and understanding of adaptive behavior
The adaptations were clear	0.87
The adaptations were natural	0.80
I understand why adaptations were made	0.73
The adaptations were not annoying	0.67
The adaptations helped me	0.63
Eigenvalue	**2.77**

Factor-scores were saved using the regression method. The measure is normally distributed based on Kolmogorov-Smirnov and Shapiro-Wilk tests.

Concluding, the analysis provided a normally distributed measure of 'acceptance and understanding of adaptive behavior', based on factor scores as defined by the loadings displayed above.

Results of the 2nd experiment

This chapter describes the results of the second experiment. This experiment directly tests the hypothesis that adapting the system to the users' domain knowledge and commitment increases their satisfaction, provided that the system explains its adaptive behavior.

Specifically, the results indicate that adaptiveness without explanation actually reduces understandability, satisfaction and perceived usefulness, but that adaptiveness with explanation makes users more satisfied with the system and increases their perceived usefulness.

Contrary to our initial expectations, the results also show that a system with agent-based explanations is less satisfying than a system with neutral explanations.

Observed reactions to adaptiveness

What adaptations were made, and how participants reacted to them

Before the results of the second experiment are presented, this section first describes the adaptive behavior that occurred in the experiment, and the users' reactions to the adaptations.

User models

First of all, we found that the level of domain knowledge measured with the 4 questionnaire items was positively correlated ($r = 0.173$, $p < .05$) to the average user model level of this characteristic. This correlation is similar to the correlation found for the process data relations in experiment 1. It is however rather small, so we had to expect disappointing results of adaptiveness due to a lack of accuracy in the user model[25].

The average value of domain knowledge was 0.15 and ranged from -0.19 to 0.68. The average value of commitment was 0.16 and ranged from -0.36 to 0.72. This either means that we had many users with a high level of domain knowledge or commitment, or that our user model overestimated the value of these user characteristics.

Adaptations

Another insight in the workings of the user model is provided by an analysis of number of adaptations that were made by the system, and how many users corrected these adaptations. The results of this analysis are provided in Table 17.

Table 17: Number of adaptations made by the system, and switches made by the user (total number of participants: 132)

Adaptation / switch	# of pps adapted	# of pps immediately switching back	Total # of pps switching manually
PE method	11	5+3[26]	68
Information detail level	22	0	57
Total savings display	19	3	74
Sorting	71	12	43

Preference elicitation methods as a predictor

In experiment 1, 'preference elicitation method' was a dichotomous variable, because it was assigned randomly to each participant before the experiment and participants were not able to change the method presented to them. In experiment 2, however, the users as well as the

[25] We could not include a similar analysis for commitment, as the 4 questionnaire items measuring this characteristic did not result in a stable factor solution.

[26] An additional 3 participants did not switch back, but showed their rejection of the adaptation in another way: 2 participants did stopped using the preference elicitation and did not turn back to it; the other participant stopped the experiment.

system were able to switch the preference elicitation method during the interaction (see Table 17). Therefore, the measure for 'preference elicitation method' is in this experiment expressed as the fraction of the clicks[27] in which attribute-based preference elicitation was used, hence the variable name 'fraction attribute-based PE'. In other words, if the user started out in the attribute-based preference elicitation method, but the user or the system switched to case-based preference elicitation at the end of the interaction – e.g. after 80% of the clicks of this user – the value for 'fraction attribute-based PE' would be 0.80.

As we realize that results concerning this variable are hard to understand, especially in interactions with other variables, we provide plots with each of our analyses that display the results in a more intuitive way. The plots display the PE-method as a dichotomous variable (in line with experiment 1) that is obtained by performing a split at the 0.50 level. The plots are thus obtained by running a *different* analysis than the ones displayed in the tables; one that has reduced power (due to the dichotomization of the PE-method variable), but enhances the interpretation, as it clearly shows the difference between participants mainly using attribute-based PE and participants mainly using the case-based PE when it comes to the effect of the system type.

The effect of adaptiveness and explanations

On satisfaction with system, perceived usefulness, understandability, and acceptance/understanding of adaptive behavior

In order to test hypotheses H8, H9 and H10, we performed linear regressions using system type (static, adaptive no explanations, adaptive with explanations, adaptive with agent-based explanations), domain knowledge and preference elicitation method as predictors[28], and satisfaction with the system, perceived usefulness, understandability, and acceptance and understanding of the adaptive behavior as dependent variables. In the analyses, each of the adaptive systems was separately compared to the static system, which provided a baseline result. Linear regressions were used because the sum scale of satisfaction and the factor scores of the other measures were constructed as an interval scale. Predictors were either nominal (system type), interval (domain knowledge and commitment), or ratio (fraction attribute-based PE).

Predicting 'satisfaction with the system'

Table 18 and Figure 14 present the results of the regression on satisfaction. H8 and H9 predicted an increase of satisfaction for users of the adaptive system with explanations and the

[27] We used the fraction of the clicks and not the fraction of the time here, because this gives us a more stable measure, as we could not prevent users from interrupting their interaction for a certain period of time.

[28] Commitment could not be used as a predictor in this experiment, as the 4 questionnaire items measuring this characteristic did not result in a stable factor solution.

adaptive system with agent-based explanations respectively (compared to participants using the 'static' system). The small positive significant effect of the interaction of the 'adaptive with explanations' system with 'fraction attribute-based PE' partially confirmed H8. Specifically, this effect shows that users of the adaptive system with explanations rated the system as more satisfying if they used the attribute-based PE method more extensively (see Figure 14). In other words, the beneficial effect of adaptiveness with explanations holds only for participants mainly using the attribute-based PE method.

No evidence was found to support H9, as the adaptive system with agent-based explanation showed no significant improvements over the static system, nor did the interaction with any of the other variables.

H10 predicted that a lack of explanations would severely reduce satisfaction compared to the static system. This would show as a negative significant effect of the 'adaptive, no explanations' system type in the current analysis. This effect is suggested by a negative estimate (B = -4.510), but it was not significant ($p > 0.10$).

Table 18: Predicting satisfaction (adjusted $R^2 = .126$)

	Estimate	Std. Error	t	Partial n^2
Intercept	24.873	2.055	12.104***	0.560
System type[29]				
Adaptive, no explanations	-4.510	2.734	-1.650	0.023
Adaptive with explanations	-3.234	2.797	-1.156	0.011
Adaptive with agent	-1.648	3.142	-0.524	0.002
Domain knowledge	-0.273	2.332	-0.117	0.000
Fraction attribute-based PE	2.063	2.751	0.750	0.005
System type * Domain knowledge				
Adaptive, no explanations	-0.085	3.369	-0.025	0.000
Adaptive with explanations	-3.298	2.818	-1.170	0.012
Adaptive with agent	3.632	3.754	0.968	0.008
System type * Fraction attribute-based PE				
Adaptive, no explanations	2.768	3.803	0.728	0.005
Adaptive with explanations	8.280	3.889	2.129*	0.038
Adaptive with agent	2.287	4.142	0.552	0.003
Domain knowledge * Fraction attribute-based PE	-1.252	3.408	-0.367	0.001
System type * Domain knowledge * Fraction attribute-based PE				
Adaptive, no explanations	2.230	4.645	0.480	0.002
Adaptive with explanations	7.239	4.363	1.659	0.023
Adaptive with agent	-4.080	5.108	-0.799	0.006

*$p < .05$ **$p < .01$ ***$p < .001$

[29] Using the 'static' system as the reference category

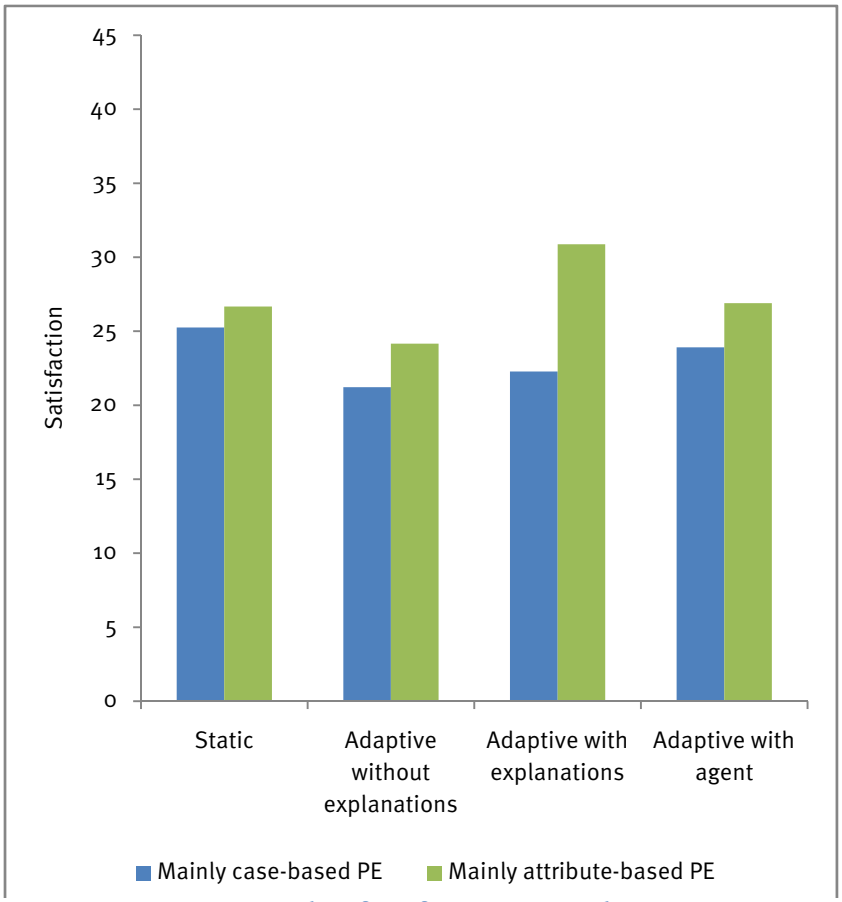


Figure 14: Value of satisfaction across conditions

Predicting perceived usefulness

The results of the regression on perceived usefulness are displayed in Table 19 and Figure 15 below. H8 and H9 not only predicted that the two 'explanation' and 'adaptive with agent-based explanations' levels of adaptiveness would result in an increased satisfaction, but also that participants in these conditions would perceive the system as more useful (compared to participants using the 'static' system). Again, we confirmed H8 with a small positive significant effect of the interaction of the 'adaptive with explanations' system type with 'fraction attribute-based PE'. The increase in usefulness for the 'explanation' system type was thus again contingent on the amount of time participants spent using attribute-based PE. Specifically, the benefit of the explanations was higher when participants used the attribute-based PE-method more extensively.

Again, no evidence was found to support H9, as the adaptive system with agent-based explanations showed no significant increase in perceived usefulness over the static system.

H10 predicted – besides a decrease in satisfaction – also a decrease in perceived usefulness for the adaptive system without explanations. We confirmed this hypothesis with a small negative significant effect of the 'adaptive, no explanations' system type. However, the results of this analysis suggest that this effect is reduced when participants make more extensive use of the

attribute-based PE-method: the main effect of the 'adaptive, no explanations' system type is a decrease in perceived usefulness (B = -0.788), but the interaction effect of the 'adaptive, no explanations' system type with the 'fraction attribute-based PE' has a similar increase in perceived usefulness[30] (B = +0.755). This means that for participants that mainly use the case-based PE-method, the negative effect of adaptiveness without explanation is reduced to almost zero. This point is illustrated more clearly in Figure 15, where the difference between the leftmost two green bars is smaller than the difference between the leftmost two blue bars.

Table 19: Predicting perceived usefulness (R2 = .192)

	Estimate	Std. Error	t	Partial n^2
Intercept	0.253	0.249	1.014	0.009
System type				
Adaptive, no explanations	-0.788	0.331	-2.377*	0.047
Adaptive with explanations	-0.523	0.339	-1.544	0.020
Adaptive with agent	-0.310	0.381	-0.815	0.006
Domain knowledge	-0.475	0.283	-1.682	0.024
Fraction attribute-based PE	-0.332	0.333	-0.997	0.009
System type * Domain knowledge				
Adaptive, no explanations	0.275	0.408	0.675	0.004
Adaptive with explanations	-0.186	0.342	-0.546	0.003
Adaptive with agent	0.523	0.455	1.150	0.011
System type * Fraction attribute-based PE				
Adaptive, no explanations	0.755	0.461	1.638	0.023
Adaptive with explanations	1.254	0.471	2.662**	0.058
Adaptive with agent	0.321	0.502	0.639	0.004
Domain knowledge * Fraction attribute-based PE	-0.448	0.413	-1.084	0.010
System type * Domain knowledge * Fraction attribute-based PE				
Adaptive, no explanations	0.680	0.563	1.208	0.013
Adaptive with explanations	1.046	0.529	1.978	0.033
Adaptive with agent	0.120	0.619	0.193	0.000

*$p < .05$ **$p < .01$ ***$p < .001$

[30] even though this effect is not significant ($p > .10$)

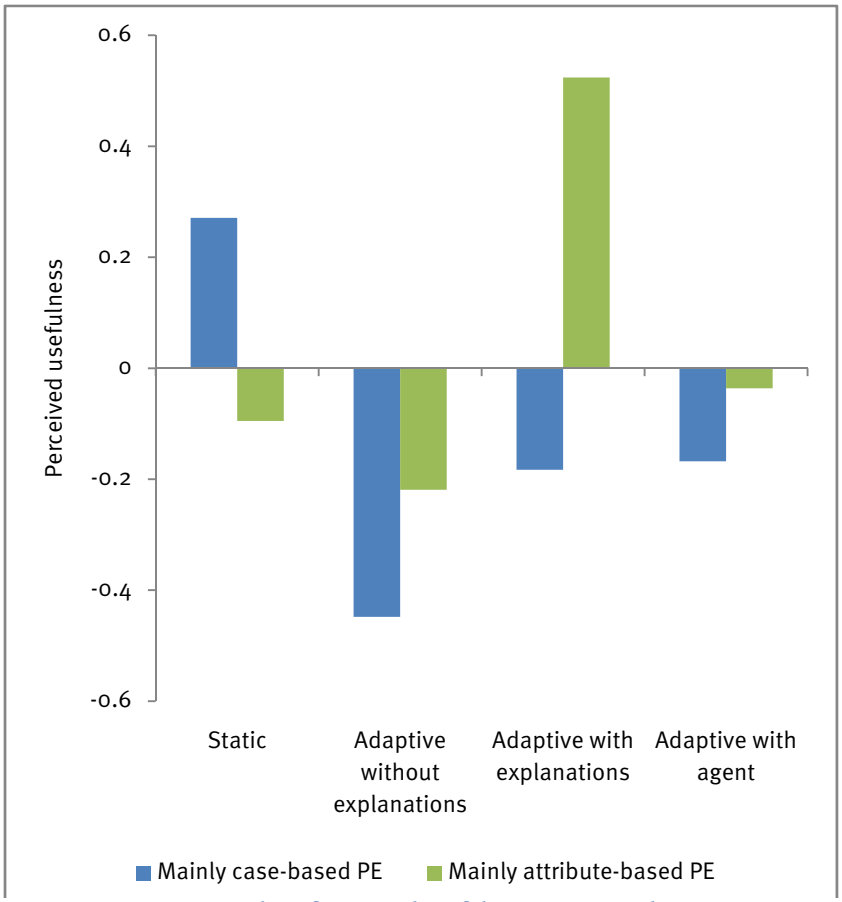


Figure 15: Value of perceived usefulness across conditions

Predicting understandability

The results of the regression on understandability are displayed in Table 20 below.

Predicting understandability, we only found weak evidence for hypothesis H10; an adaptive system without explanation was less understandable than the baseline system, but this effect was only marginally significant ($p < .10$).

Running the same analysis without the 'system type' variable provided a large significant effect of 'fraction attribute-based PE' ($p < .001$, partial $\eta^2 = .145$). In other words, just like in experiment 1, the attribute-based PE method is more understandable than the case-based PE method.

Table 20: Predicting understandability (adjusted R^2 = .118)

	Estimate	Std. Error	t	Partial n^2
Intercept	-0.040	0.263	-0.153	0.000
System type				
Adaptive, no explanations	-0.618	0.349	-1.769	0.026
Adaptive with explanations	-0.521	0.357	-1.457	0.018
Adaptive with agent	-0.444	0.401	-1.105	0.011
Domain knowledge	0.240	0.298	0.804	0.006
Fraction attribute-based PE	0.490	0.352	1.394	0.017
System type * Domain knowledge				
Adaptive, no explanations	-0.420	0.430	-0.976	0.008
Adaptive with explanations	-0.588	0.360	-1.632	0.023
Adaptive with agent	0.233	0.480	0.486	0.002
System type * Fraction attribute-based PE				
Adaptive, no explanations	0.313	0.486	0.643	0.004
Adaptive with explanations	0.685	0.497	1.378	0.016
Adaptive with agent	0.189	0.529	0.358	0.001
Domain knowledge * Fraction attribute-based PE	-0.194	0.435	-0.446	0.002
System type * Domain knowledge * Fraction attribute-based PE				
Adaptive, no explanations	0.646	0.594	1.089	0.010
Adaptive with explanations	0.931	0.558	1.671	0.024
Adaptive with agent	-0.475	0.653	-0.728	0.005

*$p < .05$ **$p < .01$ ***$p < .001$

Predicting 'acceptance and understanding of adaptive behavior'

H10 predicted a decrease in the acceptance and understanding of the adaptive behavior for participants in using the adaptive system without explanation, compared to participants using the adaptive system with explanations or with agent-based explanations. In order to test this hypothesis, we performed a regression on 'acceptance and understanding of the adaptive behavior' with the system type as a predictor[31]. We also included the number of adaptations made by the system as a predictor, as we reasoned that the amount of adaptiveness experienced by the participants might also increase or decrease their acceptance and understanding of the adaptive behavior. The results of this analysis are displayed in Table 21 and [figure] below.

[31] Note that this measure was only taken for participants that actually experienced adaptations. The analysis therefore does not include the 'static' system, as this system made no adaptations.

Table 21: Predicting acceptance and understanding of adaptive behavior (adjusted R^2 = .087)

	Estimate	Std. Error	t	Partial n^2
Intercept	-1.417	0.490	-2.893**	0.119
System type[32]				
Adaptive with explanations	1.651	0.541	3.050**	0.130
Adaptive with agent	1.394	0.571	2.441*	0.088
Number of adaptations	0.371	0.136	2.724**	0.107
Domain Knowledge	-0.079	0.492	-0.160	0.000
System type * Domain knowledge				
Adaptive with explanations	-0.728	0.733	-0.993	0.016
Adaptive with agent	0.032	0.571	0.055	0.000
System type * Number of adaptations				
Adaptive with explanations	-0.371	0.141	-2.634*	0.101
Adaptive with agent	-0.348	0.145	-2.406*	0.085
Number of adaptations * Domain Knowledge	-0.069	0.128	-0.537	0.005
System type * Number of adaptations * Domain knowledge				
Adaptive with explanations	0.383	0.217	1.766	0.048
Adaptive with agent	0.015	0.135	0.110	0.000

*$p < .05$ **$p < .01$ ***$p < .001$

We confirmed H10 with the two medium-sized significant main effects of the 'adaptive with explanations' en 'adaptive with agent-based explanations' system types. This means that participants in these conditions showed more acceptance and understanding of the adaptive behavior than participants in the 'adaptive, no explanations' condition.

The medium-sized significant positive main effect of 'number of adaptations' shows that more adaptations make the adaptive behavior more acceptable and understandable. Moreover, the medium-sized positive significant interaction effects of the 'adaptive with explanations' (B = 1.651) and 'adaptive with agent' (B = 1.394) conditions with the number of adaptations indicate that the positive effect of explanations is smaller the more adaptations the system makes: the effect is 0.371 (for 'adaptive with explanations') and 0.348 (for 'adaptive with agent') lower per adaptation. This can be explained by reasoning that participants may get used to the adaptive behavior when they experience it more frequently.

[32] Using the adaptive system without explanations as the reference category.

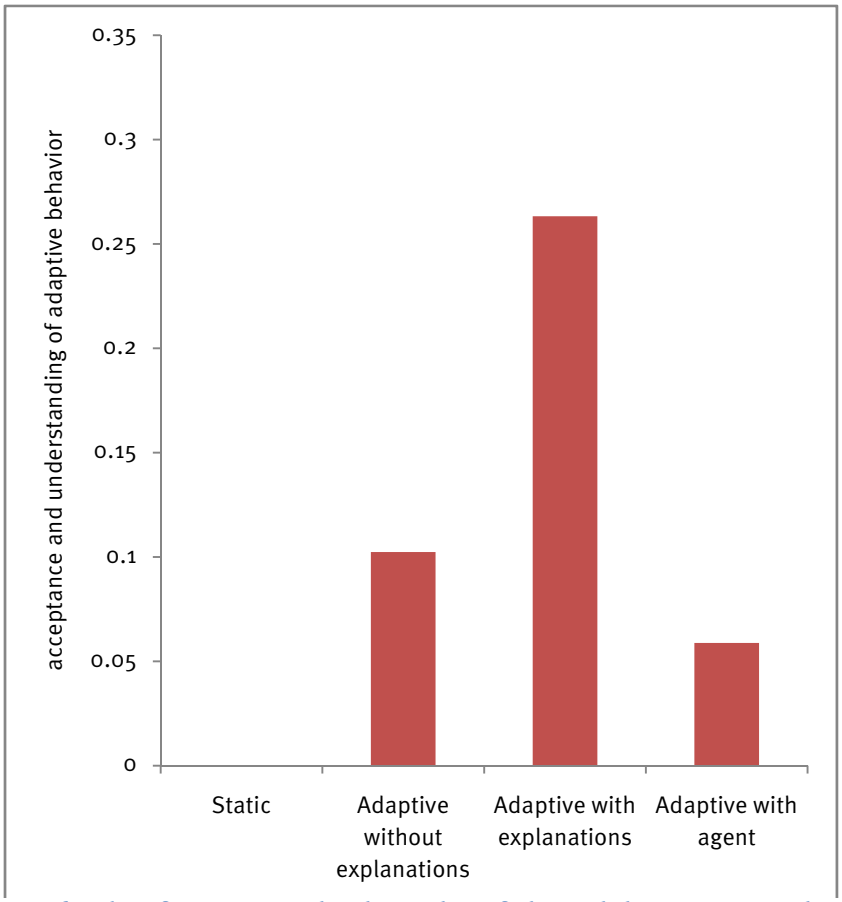


Figure 16: Value of acceptance and understanding of adaptive behavior across conditions

The effect of agent-based explanations

On acceptance/understanding of the adaptive behavior and perceived personal help

H11 predicted that participants using the adaptive system with agent-based explanations contend that the system provides more personal help, and that they are more willing to accept the adaptive behavior than people using the other adaptive systems. In order to test this hypothesis, we analyzed the potential increase in the 'acceptance and understanding of the adaptive behavior' and 'perceived personal help' in the 'adaptive with agent-based explanations' condition.

Predicting 'acceptance and understanding of adaptive behavior'

If H11 is correct, it would result in a significant increase in 'acceptance and understanding of the adaptive behavior' for the 'adaptive with agent-based explanations' system type compared to the 'adaptive with explanations' system type.

In the previous paragraph, we already discussed the result of the analysis of 'acceptance and understanding of the adaptive behavior', and found that the 'adaptive with explanations' and 'agents' condition showed an increase over the 'static' condition. As can be seen in Table 21, however, the people using the 'adaptive with agents' system actually show a *decrease* over the 'adaptive with explanations' system: The estimate of 'adaptive with agent' (B = 1.651) is lower than the estimate of 'adaptive with explanations' (B = 1.394). This means that the analysis did not support H11.

Predicting 'perceived personal help'

H11 also predicts an increase in the 'perceived personal help' for participants using the 'adaptive with agent-based explanations' system. This hypothesis was tested with a linear regression using system type, domain knowledge and preference elicitation method as predictors. Results of this analysis are shown in Table 22 and Figure 17.

Table 22: Predicting perceived personal help (adjusted R^2 = .173)

	Estimate	Std. Error	t	Partial n^2
Intercept	0.057	0.260	0.221	0.000
System type				
Adaptive, no explanations	-0.465	0.346	-1.346	0.016
Adaptive with explanations	-0.527	0.354	-1.491	0.019
Adaptive with agent	-0.307	0.397	-0.774	0.005
Domain knowledge	-0.589	0.295	-1.999*	0.034
Fraction attribute-based PE	-0.260	0.348	-0.746	0.005
System type * Domain knowledge				
Adaptive, no explanations	0.256	0.426	0.601	0.003
Adaptive with explanations	-0.070	0.356	-0.195	0.000
Adaptive with agent	0.680	0.474	1.433	0.018
System type * Fraction attribute-based PE				
Adaptive, no explanations	0.528	0.481	1.098	0.010
Adaptive with explanations	1.548	0.492	3.150**	0.079
Adaptive with agent	0.780	0.524	1.490	0.019
Domain knowledge * Fraction attribute-based PE	0.086	0.431	0.200	0.000
System type * Domain knowledge * Fraction attribute-based PE				
Adaptive, no explanations	0.385	0.587	0.655	0.004
Adaptive with explanations	0.806	0.552	1.461	0.018
Adaptive with agent	-0.497	0.646	-0.770	0.005

*$p < .05$ **$p < .01$ ***$p < .001$

The fact that this hypothesis was not supported can best be observed in Figure 17; the 'perceived personal help' was not consistently higher for the 'adaptive with agent-based

explanations' system type than the other levels. Table 22 confirms this with no significant results for the 'adaptive with agent-based explanations' condition.

We did however find a medium-sized effect of the interaction of the 'adaptive with explanations' system type with 'fraction attribute-based PE'. This means that participants that more extensively used the attribute-based PE method indicated that the adaptive system with 'generic' explanations provided more personal help.

Finally, we found a small negative significant effect of 'domain knowledge'. This means that experts were in general less likely to indicate the system to be personally helpful and adaptive than novices.

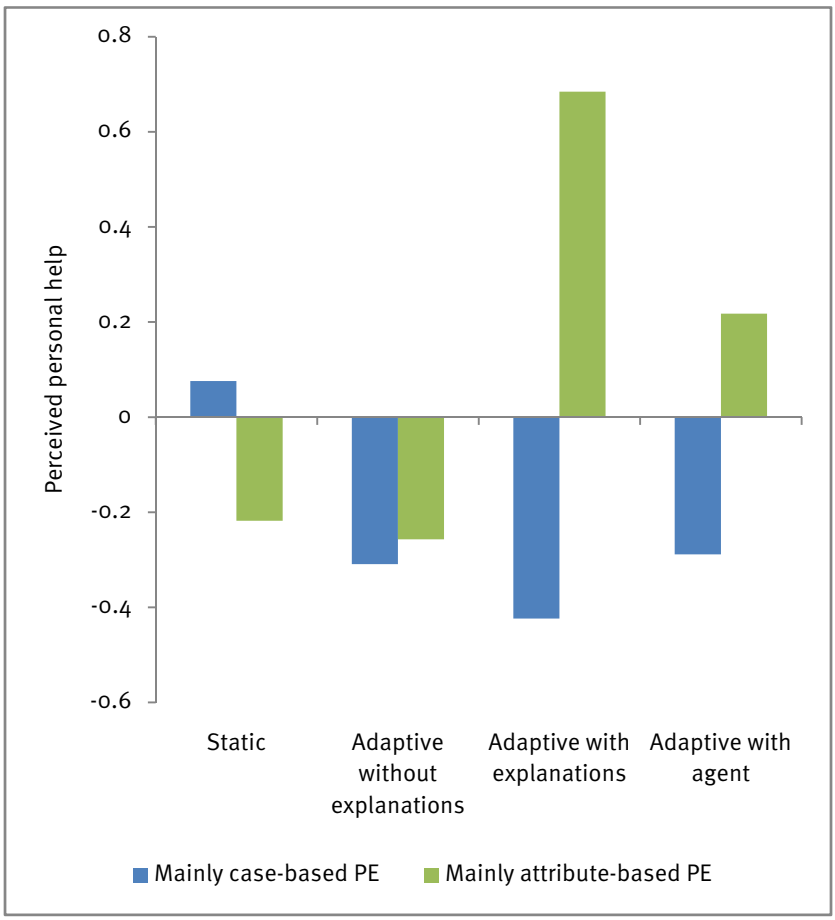


Figure 17: Value of perceived personal help across conditions

Additional observations

Predicting satisfaction with the chosen measures and total amount of energy saved

Predicting satisfaction with the chosen measures

Based on the results of experiment 1, we predicted that an effect of satisfaction with the chosen measures would be mediated by the subjective system-related measures 'satisfaction

with the system' (as in experiment 1) and 'perceived personal help' (a subjective measure introduced in experiment 2). A regression with these measures as predictors provided the results shown in Table 23 below.

We found a small significant effect of 'satisfaction with the system' and a medium-sized significant effect of 'perceived personal help' on 'satisfaction with the chosen measure'. In general, the satisfaction with a recommender system and the perceived personal help it offers can reflect on the items chosen/purchased with the system.

Table 23: Predicting satisfaction with the chosen measures (adjusted R^2 = .326)

	Estimate	Std. Error	t	Partial n^2
Intercept	-0.756	0.331	-2.287*	0.039
Satisfaction with system	0.029	0.013	2.332*	0.041
Perceived personal help	0.365	0.097	3.753***	0.099

*$p < .05$ **$p < .01$ ***$p < .001$

Predicting total amount of energy saved

As in experiment 1, we wanted to check whether our different systems would have a varying effect on the total amount of energy saved by the participants. Table 24 and Figure 18 present the results of the regression on total amount of energy saved (in kilowatt-hour).

Interestingly, the total amount of energy saved *decreased* significantly (compared to the 'static' system) when participants made more extensive use of the case-based preference elicitation method, and were interacting with the agent-based system. Note that there was no decrease (compared to the 'static' system) in amount of energy saved when participants in the 'agent' condition used the attribute-based preference elicitation method, because the positive interaction effect of the 'adaptive with agent' system with 'fraction of attribute-based PE' (B = 310.70) extinguished the negative main effect of the 'adaptive with agent' system (B = -283.64).

Table 24: Predicting total amount of energy saved (adjusted R^2 = .184)

	Estimate	Std. Error	t	Partial n^2
Intercept	335.77	77.260	4.346***	0.155
System type				
Adaptive, no explanations	-157.15	105.680	-1.487	0.021
Adaptive with explanations	-61.90	102.769	-0.602	0.004
Adaptive with agent	-283.64	120.907	-2.346*	0.051
Domain knowledge	-9.48	84.472	-0.112	0.000
Fraction attribute-based PE	121.12	102.735	1.179	0.013
System type * Domain knowledge				
Adaptive, no explanations	44.69	126.360	0.354	0.001
Adaptive with explanations	-1.55	101.608	-0.015	0.000
Adaptive with agent	76.73	136.713	0.561	0.003
System type * Fraction attribute-based PE				
Adaptive, no explanations	-78.84	144.924	-0.544	0.003
Adaptive with explanations	58.11	144.163	0.403	0.002
Adaptive with agent	310.70	156.501	1.985*	0.037
Domain knowledge * Fraction attribute-based PE	203.93	122.567	1.664	0.026
System type * Domain knowledge * Fraction attribute-based PE				
Adaptive, no explanations	-211.97	169.822	-1.248	0.015
Adaptive with explanations	-56.18	158.372	-0.355	0.001
Adaptive with agent	-213.48	184.486	-1.157	0.013

*$p < .05$ **$p < .01$ ***$p < .001$

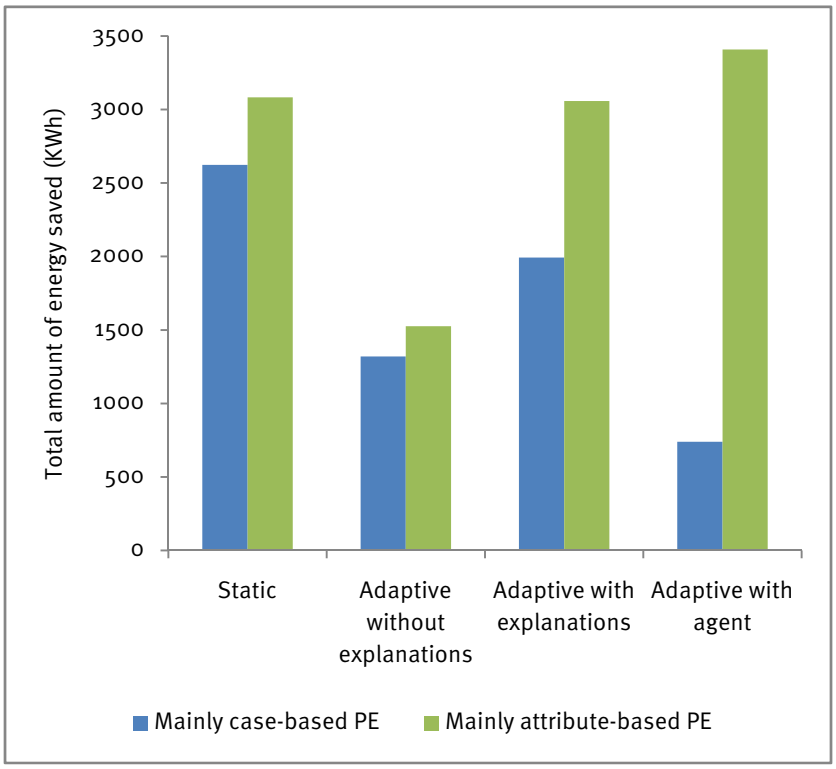


Figure 18: Total amount of energy saved across conditions

Conclusion

Adaptation works, but only with explanations; the agent is disappointing

Only two out of the four hypotheses put forward in experiment 2 were (partially) confirmed. In line with H8, participants using the adaptive system with explanations judged the system to be more satisfying and more useful, and to provide better personal help than participants using the 'static' system. This was however only true for participants that mainly used the attribute-based preference elicitation. The reason why this distinction exists remains open for further investigation in future research.

Furthermore, as predicted by H9, the adaptive system without explanations was judged to be less understandable and less satisfying than the 'static' variant, but only slightly. The adaptations were also less understandable and acceptable in this condition. Moreover, this system was judged to be less useful, but only by participants mainly using the case-based preference elicitation. Again, the reason for this distinction remains an open issue.

Finally, contrary to what was predicted in H10, the positive effect of explanations was only found for the system providing 'generic' explanations, and not for the system with agent-based explanations. The increase in perceived personal help and acceptance of the adaptive behavior, predicted in H11, was also not confirmed.

Concluding, the adaptive system with generic explanations provided the best results, albeit only for participants that mainly used the attribute-based preference elicitation. The adaptive system without expectations performed the worst. Unexpectedly, the agent-based system did not improve user satisfaction, and generally provided worse results than the adaptive system with generic explanations.

Discussion

The correlation of the user model value of domain knowledge with the survey measure was significant but quite low, and this may explain why the adaptive system did not provide universally positive results. Moreover, only few adaptations were made, and to some types of adaptations quite a few corrections were made. This indicates the imperfection of our user modeling: the system was not able to adapt to every participant, and some participants disliked the adaptations and consequently corrected them.

In general, participants mainly using the attribute-based preference elicitation method evaluated the adaptive behavior more positive. They showed no decrease compared to the 'static' system for the 'adaptive, no explanations' system, and an increase for the 'adaptive with explanations' system. One could argue that this may be due to the higher understandability of this preference elicitation method. Specifically, since the case-based preference elicitation method was already less understandable than the attribute-based preference elicitation method, the extra confusion introduced by the adaptive behavior may have caused the total complexity of the system to reach beyond the users' cognitive capacity.

Conclusion and discussion

This chapter returns to the central thesis argument, which argued that – compared to a 'traditional' recommender system – an adaptive recommender system with agent-based explanations will have positive effects on the satisfaction and the choices made by people using the system.

While the experiments described in the previous chapters use the case of choosing energy-saving measures as a typical case in which different levels of domain knowledge and choice goals warrant the success of an adaptive system, this chapter extends the case to recommender systems in general, be it advisory systems, web shops, or any other application in which the utility of choice options are expressed with attribute values.

The chapter concludes by providing advice and future opportunities for the research, design and implementation of such systems.

Findings of the current thesis

Evaluating our recommender system for energy-saving measures

A comprehensive system for choosing energy-saving measures

Although not directly relevant for the scientific implications of this thesis, we would like to acknowledge that our choice for the domain of energy-saving behavior as a use-case for our research has given it an important ideological twist. Saving energy is incredibly relevant in today's society, and the results of this thesis show that recommender systems may play an important role in the collective effort to save energy.

Thousands of energy-saving measures exist, but they are spread out across hundreds of websites and brochures. We are arguably the first to identify a choice problem in this domain by acknowledging that one cannot implement all these measures at once. Describing energy-saving measures in terms of selected attributes and presenting them in a MAUT-based recommender system may significantly help people choosing the energy-saving measures that fit their personal needs and preferences.

Tailoring to domain knowledge and commitment

Based on the literature on energy-saving behavior (e.g. Parnell & Popovic Larsen, 2005) and discussions with an energy consultant, we identified two main personal characteristics: ecological knowledge and ecological commitment. Based on decision theoretical principles (e.g. Alba & Hutchinson, 1987) we argued that people with different levels of knowledge and commitment might have different requirements when it comes to a recommender system.

Indeed, the first experiment provided evidence that experts and novices differ in which preference elicitation they found more satisfying and useful. We did not find an increase in total energy savings in the 'matched' conditions.

We also found that people with different levels of commitment framed their choice problem in different ways: committed individuals had an 'environmental benefits frame' while less-committed individuals had a 'personal benefits frame'.

The adaptive system

We recognized that measuring domain knowledge and commitment with lengthy questionnaires would be inconvenient in real life implementations of a recommender system. Therefore, we developed an adaptive system that measures domain knowledge and commitment based on process data predictors. Based on existing research on adaptive systems (e.g. Pazzani & Billsus, 2002), we argued that the adaptive system could be confusing to the users, and therefore fitted our system with explanations of the adaptive behavior.

Indeed, the second experiment provided evidence that users of an adaptive system with (neutral) explanations are in certain circumstances more satisfied with the system than users of the static version. Such a system was also perceived to be more useful and to provide more

personal help. As expected, users judged the system without explanations to be less understandable, less satisfying, and less useful than the static system.

Based on literature on agent-based interaction (e.g. Knijnenburg & Willemsen, 2008), we argued that a human-like agent explaining the adaptive behavior could potentially be more acceptable and satisfying, since a human-like appearance might intuitively imply a more flexible and intelligent interaction style. On the other hand, we acknowledged that users could potentially overestimate the capabilities of the agent, which would reduce their satisfaction and understandability.

The results of the second experiment suggest that this second consideration might be right: contrary to the system with 'neutral' explanations, the system with agent-based explanations was not perceived to be more satisfying and useful than the static version. People using the interface with the agent also did not show a higher acceptance of the adaptive behavior.

Limitations and future work

Our results show that the adaptiveness is not under all circumstances evaluated more positively. Specifically, the participants mainly using the case-based preference elicitation method rated the adaptive systems as less satisfying and less useful than the participants using the attribute-based preference elicitation method. Our intuitive explanation is that the already decreased understandability of the former preference elicitation method intensified the confusion caused by the adaptiveness. This result is however still open for interpretation, and could inspire future work on adaptive recommendation systems. For example, one could study understandability at a finer level of granularity to find out exactly which interface element causes the confusion, and whether two confusing elements may intensify each others' negative effect on user satisfaction.

The results also indicated that the user model measured the user characteristics imprecisely and this could have also caused a decrease of the benefit of adaptiveness. Repeating experiment 2 with a better user model may provide better insights in the effects of adaptiveness. Such a user model could be based on a careful analysis of the process data gathered in experiment 2, and could use novel approaches like asking questions during the interaction that would improve the certainty of the modeled values.

Moreover, it is unclear how important the adequate measurement of user characteristics really is. The relation between measurement precision and the subjective evaluation of adaptiveness does not have to be a linear one; there might for instance be a sharp drop-off at a certain level. Future studies could investigate the effect of user model measurement precision on the evaluation of adaptive recommender systems.

Adaptiveness and agent-based explanations

How decision-theoretical principles and interface design can inform the research on recommender systems

Adaptiveness, based on decision theory

Taking the liberty to generalize our findings, we have demonstrated that the well-documented individual differences in decision behavior call for a tailored approach in recommender system design. Under certain circumstances, an adaptive system increases user satisfaction, provided that the system also explains the adaptive behavior.

Although the current approach to adaptiveness is conservative compared to the technologically more advanced approach used by Hauser et al (2009), we contend that a solid theoretical foundation for the adaptive behavior reduces the risk of 'over-automating' the system. Although technological advances provide an exciting field of research, no less scientific merit can be found in a more in-depth analysis of adaptive behavior and its consequences.

Beware of the agents

A similar remark can be made about the use of human-like agents. Although a substantial number of researchers has incorporated a human-like agent in their recommender system (e.g. Bickmore & Cassell, 2001; Spiekermann, 2001; Pazzani & Billsus, 2002; Abbattista et al., 2002; Semeraro et al., 2008), none of them has tested their system explicitly against a version without the agent.

In an earlier study (Knijnenburg & Willemsen, 2008) we discovered that users expect certain human-like capabilities of a system that employs the agent metaphor, and act accordingly. The idea that adaptiveness is such a human-like capability potentially warrants the effectiveness of employing the agent metaphor in an adaptive recommender system. However, our earlier study also found that the agent-instilled expectations can lead to catastrophic usability decreases when they are not properly matched by the actual functionality of the system. The disappointing results of the agent-based explanations in our second experiment suggest that the latter may be true in our recommender system. As these findings also cast doubt on the appropriateness of the agent metaphor in the aforementioned studies, we contend that it is always advisable to test an agent-based system against its 'agentless' counterpart.

The merit of user studies

The advent of adaptive systems, agent-based interaction and advanced recommendation algorithms should not shift the focus in recommender systems research too far towards technological aspects. Minimizing root mean square prediction errors and running simulation studies are both valuable endeavors, but since in the end real human beings have to use the systems, they should be the main focus of our research (Xiao & Benbasat, 2007). This is evident from the current thesis: although experiment 1 warranted the *potential* merit of an

adaptive system, experiment 2 showed that such an adaptive approach is subject to several intricate phenomena, like the effect of proper explanations. User studies are therefore an intricate part of recommender systems research.

User-focused recommender systems research

Concluding, this thesis is an attempt at a user-focused study of recommender systems. Principles of decision theory, psychology and interface design were used to make improvements to a recommender system, and user-testing was employed to confirm the benefit of these improvements. We thus acknowledge that a multi-disciplinary approach can improve the usability of recommender systems, thereby increasing user satisfaction, and we believe that a sound understanding of the aforementioned fields will play a key role in the future research and development of recommender systems.

Acknowledgements

Thanks and praise

I would like to thank my supervisor Martijn Willemsen for his collaboration, supervision and ongoing support during the project; my second supervisor Benedict Dellaert for his help with the utility model calibration experiment, his relevant pointers to related literature, and the final review of my thesis; Steven Langerwerf for his extensive advice in the design, development and optimization of the Web Recommender System, for hosting the experiments, and for his perspectives on the technological context of adaptive recommender systems; the rest of the academic staff of the Human-Technology Interaction group for their interest and feedback; Evelien Matthijssen for her input on personal characteristics in energy-saving and a preliminary list of energy-saving measures; Roelof Lochmans and Dylan Schouten for their help with the construction of a list of energy-saving measures, and for their help setting up experiment 1 and the precursory pre-tests; Jeroen van Agt and Marcel van der Steen for distributing the experiments online; and finally my parents and my girlfriend for much-needed mental support.

Works Cited

Abbattista, F., Degemmis, M., Licchelli, O., Lops, P., Semeraro, G., & Zambetta, F. (2002). Improving the usability of an e-commerce web site through personalization. *Recommendation and Personalization in Ecommerce, 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems* (pp. 20-29). Malaga, Spain: Springer.

Abbattista, F., Lops, P., Semeraro, G., Andersen, V., & Andersen, H. K. (2002). Evaluating virtual agents for e-commerce. *Workshop on embodied conversational agents*. Bologna, Italy: vhml.org.

Aberg, J., & Shahmehri, N. (2000). The role of human Web assistants in e-commerce: an analysis and a usability study. *Internet Research: Electronic Networking Applications and Policy , 10* (2), 114-125.

Alba, J. W., & Hutchinson, J. W. (1987). Dimensions of Consumer Expertise. *Journal of Consumer Research , 13*, 411-454.

Alba, J., Lynch, J., Weitz, B., Janiszewski, C., Lutz, R., Sawyer, A., et al. (1997). Interactive Home Shopping: Consumer, Retailer, and Manufacturer Incentives to Participate in Electronic Marketplaces. *Journal of Marketing , 61*, 38-53.

Andersen, V., & Andersen, H. H. (2002). *Evaluation of the COGITO system*. Roskilde, Denmark: Riso National Laboratory.

Andersen, V., Hansen, C. B., & Andersen, H. H. (2001). *Evaluation of Agents and Study of End-user needs and behaviour for E-commerce*. Roskilde, Denmark: Riso National Laboratory.

Angehrn, A. A. (1993). Computers that Criticize You: Stimulus-Based Decision Support Systems. *Interfaces , 23* (3), 3-16.

Banfi, S., Farsi, M., Massimo, F., & Jakob, M. (2008). Willingness to pay for energy-saving measures in residential buildings. *Energy Economics , 30*, 503-516.

Barr, S., Gilg, A. W., & Nicholas, F. (2005). The household energy gap: examining the divide between habitual- and purchase-related conservation behaviours. *Energy Policy , 33*, 1425-1444.

Bettman, J. R., & Park, C. W. (1980). Effects of Prior Knowledge, Exposure, and Phase of the Choice Process on Consumer Decision Processes: A Protocol analysis. *Journal of Consumer Research , 7*, 234-248.

Bettman, J. R., Luce, M. F., & Payne, J. W. (1998). Constructive Consumer Choice Processes. *Journal of Consumer Research , 25* (3), 187-217.

Bickmore, T., & Cassell, J. (2001). Relational Agents: A Model and Implementation of Building User Trust. *Proceedings of the Conference of Human Factors in Computing Systems (SIGCHI 2001)* (pp. 396-403). New York: ACM Press.

Burke, R. D., Hammond, K. J., & Young, B. C. (1997). The FindMe Approach to Assisted Browsing. *IEEE Expert , 12* (4), 32-40.

Chai, J., Horvath, V., Nicolov, N., Stys, M., Kambhatla, N., Zadrozny, W., et al. (2002). Natural Language Assistant: A Dialog System for Online Product Recommendation. *AI Magazine , 23* (2), 63-76.

Chernev, A. (2003). When More Is Less and Less Is More: The Role of Ideal Point Availability and Assortment in Consumer Choice. *Journal of Consumer Research , 30*, 170-183.

Cheung, C. M., Chan, G. W., & Limayem, M. (2005). A Critical Review of Online Consumer Behavior: Empirical Research. *Journal of Electronic Commerce in Organizations , 3* (4), 1-19.

Clark, A. (2003). *Natural-Born Cyborgs; Minds, Technologies, and the Future of Human Intelligence*. Oxford: Oxford University Press.

Cook, J., & Salvendy, G. (1989). Perception of computer dialogue personality: An exploratory study. *International Journal of Man- Machine Studies, 31* , 717- 728.

Coupey, E., Irwin, J. R., & Payne, J. W. (1998). Product Category Familiarity and Preference Construction. *Journal of Consumer Research , 24*, 459-468.

Cox, J., & Dale, B. G. (2001). Service quality and e-commerce: an exploratory analysis. *Managing Service Quality , 11* (2), 121-131.

Darby, S. (2003). Making sense of energy advice. *ECEEE 2003 Summer Study proceedings* (pp. 1217-1226). Saint-Raphaël, France: European Council for an energy efficient economy.

Dietz, T., Stern, P. C., & Guagnano, G. A. (1998). Social Structural and Social Psychological Bases of Environmental Concern. *Environment and Behavior , 30* (4), 450-471.

Dunlap, R. E., Van Liere, K. D., Mertig, A. G., & Jones, R. E. (2000). Measuring Endorsement of the New Ecological Paradigm: A Revised NEP Scale. *Journal of Social Issues , 56* (3), 425-442.

Farsi, M. (2008). *Risk-Aversion and Willingness to Pay for Energy Efficient Systems in Rental Apartments*. Swiss Federal Institutes of Technology, Centre for Energy Policy and Economics. Zürich, Switzerland: CEPE.

Flynn, L. R., & Goldsmith, R. E. (1999). A Short, Reliable Measure of Subjective Knowledge. *Journal of Business Research , 46*, 57-66.

Gutman, J. (1982). A Means-End Chain Model Based on Consumer Categorization Processes. *The Journal of Marketing , 46* (2), 60-72.

Guttman, R. H. (1998). *Merchant Differentiation through Integrative Negotiation in Agent-mediated Electronic Commerce*. Cambridge, MA, USA: Massachusetts Institute of Technology.

Guttman, R. H., & Maes, P. (1998). Agent-Mediated Integrative Negotiation for Retail Electronic Commerce. *First International Workshop on Agent Mediated Electronic Trading AMET-98* (pp. 70-90). Minneapolis, MN, USA: Springer Berlin / Heidelberg.

Guttman, R. H., Moukas, A. G., & Maes, P. (1998). Agent-mediated Electronic Commerce: A Survey. *The Knowledge Engineering Review , 13* (2), 147-159.

Haubl, G., & Trifts, V. (2000). Consumer Decision Making in Online Shopping Environments: The Effects of Interactive Decision Aids. *Marketing Science , 19* (1), 4-21.

Hauser, J. R., Urban, G. L., Liberali, G., & Braun, M. (2009). Website Morphing. *Marketing Science , 28*, 202-223.

Holzwarth, M., Janiszewski, C., & Neumann, M. M. (2006). The Influence of Avatars on Online Consumer Shopping Behavior. *Journal of Marketing , 70* (10), 19-36.

Höök, K. (1998). Evaluating the utility and usability of an adaptive hypermedia system. *Knowledge-Based Systems , 10*, 311-319.

Höök, K. (2000). Steps to take before intelligent user interfaces become real. *Interacting with Computers* , 409-426.

Jameson, A. (2002). Adaptive interfaces and agents. In J. A. Jacko, & A. Sears (Eds.), *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications* (pp. 305-330). Hillsdale, NJ, USA: L. Erlbaum Associates Inc.

Kaiser, F. G. (1998). A general measure of ecological behavior. *Journal of Applied Social Psychology , 28*, 395-422.

Kaiser, F. G., Wölfing, S., & Fuhrer, U. (1999). Environmental Attitude and Ecological Behaviour. *Journal of Environmental Psychology , 19*, 1-19.

Keeling, K., McGoldrick, P., Beatty, S., & Macaulay, L. (2004). Face Value? Customer views of appropriate formats for embodied conversational agents (ECAs) in online retailing. *37th Hawaii International Conference on System Sciences* (pp. 178-187). Big Island, Hawaii, USA: IEEE.

Knijnenburg, B. P., & Willemsen, M. C. (2008). Inferring capabilities of intelligent agents from their external traits. *(manuscript)* .

Langerwerf, S. T. (2009). *Improving Usability through Tailored User Interfaces using Software Operation Knowledge*. Working Paper.

Laurel, B. (1990). Interface agents: Metaphors With Character. In B. Laurel (Ed.), *The Art of Human-Computer Interface Design*. Reading, MA: Addison-Wesley.

Li, J. (1999). *User Interface Agents in Electronic Commerce Applications*. M. Comp. Sc. Thesis, Concordia University, Department of Computer Science, Montreal.

Li, N., & Zhang, P. (2002). Consumer Online Shopping Attitudes and Behavior: An Assessment Of Research. *Proceedings of the Eighth Americas Conference on Information Systems* (pp. 508-517). Dallas, Texas, USA: Association for Information Systems.

Maes, P., Guttman, R. H., & Moukas, A. G. (1999). Agents That Buy and Sell. *Communications of the ACM , 42* (3), 81-91.

McBreen, H. M., & Jack, M. A. (2001). Evaluating Humanoid Synthetic Agents in E-Retail Applications. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans , 31* (5), 394-405.

McGinty, L., & Smyth, B. (2002). Comparison-Based Recommendation. In S. Craw, & A. Preece (Ed.), *Advances in Case-Based Reasoning, 6th European Conference, ECCBR* (pp. 575-589). Aberdeen, Scotland, UK: Springer Berlin / Heidelberg.

McMakin, A. H., Malone, E. L., & Lundgren, R. E. (2002). Motivating Residents to Conserve Energy without Financial Incentives. *Environment and Behavior , 34*, 848-863.

McSherry, D. (2003). Similarity and Compromise. *Case-Based Reasoning Research and Development, 5th International Conference on Case-Based Reasoning, ICCBR* (pp. 291-305). Trondheim, Norway: Springer Berlin / Heidelburg.

Olson, E. L., & Widing II, R. E. (2002). Are Interactive Decision Aids Better than Passive Decision Aids? A Comparison with Implications for Information Providers on the Internet. *Journal of Interactive Marketing , 16* (2), 22-33.

Parnell, R., & Popovic Larsen, O. (2005). Informing the Development of Domestic Energy Efficiency Initiatives - An Everyday Householder-centered Framework. *Environment and Behavior , 37*, 787-807.

Pazzani, M. J., & Billsus, D. (2002). Adaptive Web Site Agents. *Autonomous Agents and Multi-Agent Systems* , 205-218.

Poortinga, W., Steg, L., Vlek, C., & Wiersma, G. (2003). Household preferences for energy-saving measures: A conjoint analysis. *Journal of Economic Psychology , 24*, 49-64.

Pu, P. H., & Chen, L. (2005). Integrating Tradeoff Support in Product Search Tools for E-Commerce Sites. *Proceedings of the ACM Conference on Electronic Commerce* (pp. 269-278). Vancouver, British Columbia, Canada: ACM.

Pu, P. H., & Kumar, P. (2004). Evaluating Example-based Search Tools. *Proceedings of the ACM Conference on Electronic Commerce* (pp. 208-217). New York, NY: ACM.

Qiu, L., & benbasat, i. (2005). An Investigation into the Effects of Text-to-Speech Voice and 3D Avatars on the Perception of Presence and Flow of Live Help in Electronic Commerce. *ACM Transactions on Computer-Human Interaction , 12* (4), 1-27.

Qiu, L., & Benbasat, I. (forthcoming). Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems. *Journal of Management Information Systems* .

Randall, T., Terwiesch, C., & Ulrich, K. T. (2007). User Design of Customized Products. *Marketing Science , 26*, 268-280.

Schwartz, N., & Clore, G. L. (1988). How do I feel about it? The informative function of affective states. In K. Fiedler, & J. P. Forgas (Eds.), *Affect, Cognition, and Social Behavior* (pp. 44-62). Toronto: Hogrefe.

Semeraro, G., Andersen, V., Andersen, H. H., Gemmis, M. d., & Lops, P. (2008). User profiling and virtual agents: a case study on e-commerce services. *Universal Access in the Information Society , 7* (3), 179-194.

Smyth, B. (2007). Case-Based Recommendation. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The Adaptive Web* (pp. 342-376). Berlin: Springer Berlin / Heidelberg.

Smyth, B., & McClave, P. (2001). Similarity vs. Diversity. In D. W. Aha, & I. Watson (Ed.), *Case-Based Reasoning Research and Development, 4th International Conference on Case-Based Reasoning, ICCBR* (pp. 347-361). Vancouver, British Columbia, Canada: Springer Berlin / Heidelberg.

Spiekermann, S. (2001). *Online Information Search with Electronic Agents: Drivers, Impediments, and Privacy Issues*. Humboldt University. Berlin: Humboldt University.

Spiekermann, S., & Paraschiv, C. (2002). Motivating Human–Agent Interaction: Transferring Insights from Behavioral Marketing to Interface Design. *Electronic Commerce Research* , 255-285.

Stern, P. C. (2000). Toward a Coherent Theory of Environmentally Significant Behavior. *Journal of Social Issues , 56* (3), 407-424.

Van Den Poel, D., & Buckinx, W. (2005). Predicting online-purchasing behaviour. *European Journal of Operational Research , 166*, 557-575.

Van Raaij, W. F., & Verhallen, T. M. (1983). A Behavioral Model of Residential Energy Use. *Journal of Economic Psychology , 3*, 39-63.

Viappiani, P., Faltings, B., & Pu, P. (2006). Preference-based Search using Example-Critiquing with Suggestions. *Journal of Artificial Intelligence Research , 27*, 465-503.

Viappiani, P., Pu, P. H., & Faltings, B. (2007). Conversational Recommenders with Adaptive Suggestions. *Proceedings of the 2007 ACM conference on Recommender systems* (pp. 89-96). Minneapolis, MN, USA: ACM.

Wang, W., & Benbasat, I. (2007). Recommendation Agents for Electronic Commerce: Effects of Explanation Facilities on Trusting Beliefs. *Journal of Management Information Systems , 23* (4), 217-246.

Xiao, B., & Benbasat, I. (2007). E-Commerce Product Recommendation Agents: Use, Characteristics, and Impact. *MIS Quarterly , 31* (1), 137-209.

Appendices

This chapter includes all topics that are relevant to the project, but do not directly contribute to the line of reasoning in the thesis.

Appendix A is a record of the energy-saving measures used in our recommender system. Appendix B discusses the attributes of these energy-saving measures. Appendix C discusses the experiment that was conducted to calibrate the utility model of our Web Recommender System. Appendix D discusses the evolution of the design of the Web Recommender System and Appendix E describes the underlying technology used to implement it. Appendix F is a record of all pre- and post-experimental questionnaires. Appendix G discusses the implementation of adaptiveness, and how the rules for our adaptive version of the system were derived from the results of experiment 1.

List of energy-saving measures

Appendix A

A list of the energy-saving measures (in Dutch) populating the recommender system that was used in our experiments can be found in Table 25. A complete list with attribute values can be obtained from the researchers.

Table 25: A list of the energy-saving measures used in the recommender system

A++ Koel/vriescombi	Hot-fill wasmachine	Laptop in plaats van PC
A+ Koel/vriescombi	Tochtstrip op deuren aanbrengen	Groene stroom
Geen warme dingen in koelkast	Tochtstrips op ramen aanbrengen	Thermostaat 1 graad lager zetten
Koel/vrieskast ijsvrij maken	Geiser schoonmaken	Thermostaat lager bij afwezigheid
Koelkast uit bij vakantie	Koelkast op de goede plek plaatsen	Boilertemperatuur op 65 graden
Kleding luchten ipv wassen	Wasmachine volledig uitschakelen	Programmeerbare thermostaat
Drogen op waslijn	Zonneboiler	Radiatorfolie aanbrengen
A-label wasdroger met warmtepomp	Achterzijde koelkast stofvrij houden	Vervang wekker(radio) door opwind-wekker
Gasverwarmde wasdroger	Warmwater-leidingen isoleren	HR-E ketel / WKK
Lagere temperatuur wassen	Spaarlampen plaatsen	Dagelijks 20 minuten luchten
Was opsparen	LED lampen plaatsen	Brievenbus met tochtstrip
Koffie in thermoskan ipv warmhoudplaatje	Ontkalken koffiezetapparaat en/of waterkoker	Warmtewisselaar op ontluchting plaatsen
Deurdranger	Waakvlam CV doven in de zomer	Wollen deken ipv elektrisch
Koken op gas ipv elektrisch	Zonnepanelen	Tuinlampen op zonne-energie
Koken met deksel op de pan	Dakisolatie	3 minuten korter douchen
Lampen uit doen	Beeldscherm PC uitschakelen	Douche ipv bad
Senseo helemaal uitzetten	Trekbel ipv elektrische bel	Waterbesparende douchekop
Opladers ontkoppelen	Altijd gedimde lampen vervangen	Mengkraan kouder zetten
`s Avonds gordijnen/luiken sluiten	Magnetisch koelen	PC met schakeldoos uitschakelen
Telefoon met snoer	Lampje in beldrukker verwijderen	PC energiebeheer inschakelen
Dubbel glas plaatsen	Bladeren harken ipv blazen	Met de hand afwassen
Vegen ipv stofzuigen	Oven eerder uitzetten	Kaarsen
Shirts kort in de droger ipv strijken	Wasmachine ontkalken	Mini-windmolen plaatsen
Roerbakken	Vaatwasser uit na gebruik	Processor undervolten
Bewegingssensor	TFT-monitor ipv CRT	BBQ-en
Vloerisolatie	PC uitzetten bij afwezigheid	Swifferen ipv stofzuigen
Dag-nacht tarief	Thermostaat voor slapen op 14 graden	

Attributes of energy-saving measures

Appendix B

The system employed in experiment 1 and 2 holds a wide variety of energy-saving measures. The measures, 80 in total, covered a wide range, including both habitual energy-saving behaviors (like switching off the lights when you leave a room) and purchase related actions (like buying roof isolation or a more economic refrigerator) (Barr et al., 2005).

The measures were gathered from a large number of (sometimes contradictory) online sources. A list of the measures can be found in Appendix A.

Based on a series of conversations with an energy consultant and a university lecturer in sustainability studies[33], we made a careful selection of nine attributes of energy-saving measures. Their validity was checked against existing research on energy-saving behavior, and during the preliminary experiment (see Appendix B). An overview is presented in Table 26.

Table 26: Selected attributes for our energy-saving measures

Attribute	Scale	Description	Also found in
Effort once	0 to 50	The one-time effort needed to implement the measure (i.e. buying and/or installing the measure).	(Poortinga et al., 2003; McMakin et al., 2002)
Continuous effort	0 to 50	The continuous effort needed to perform the measure (i.e. repeatedly defrosting your freezer).	(Poortinga et al., 2003; McMakin et al., 2002)
Cost once	Euros	The one-time cost involved in buying the measure (i.e. purchase costs). If a non-green alternative exists, these are the *additional* purchase costs.	(Barr et al., 2005; Van Raaij & Verhallen, 1983; Poortinga et al., 2003)
Continuous costs	*Euros / year*	*The repeated (additional) costs involved in the measure (i.e. additional costs of replacing energy-saving light bulbs).*	
Euro savings	*Euros / year*	*The savings in Euros on the gas or electricity bill.*	
'real' Euro savings[34]	Euros / year	The savings in Euros minus the repeated additional costs of the measure.	(Van Raaij & Verhallen, 1983)
Kilowatt-hour savings	kWh / year	The savings in kilowatt-hours on the electricity bill, or the savings on the gas bill (in m^3 gas) converted to kWh.	(Van Raaij & Verhallen, 1983; Banfi et al., 2008)
Time before return of investment	months	The time it takes to earn back the initial spending that the measure entails.	(Van Raaij & Verhallen, 1983; McMakin et al., 2002)
Comfort	-25 to 25	The increase or decrease in comfort involved in implementing the measure (i.e. taking shorter showers decreases comfort; double glazing increases comfort through noise reduction).	(Parnell & Popovic Larsen, 2005; Barr et al., 2005; Van Raaij & Verhallen, 1983; Banfi et al., 2008)
Environmental effects	-25 to 25	The positive or negative environmental effect that the measure entails, besides the energy-savings (i.e. solar panels have a negative effect, as their production costs more energy than what they save over their lifetime).	(Poortinga et al., 2003; Banfi et al., 2008)

[33] Evelien Matthijssen from www.bespaarenergie.com and Arjan Kirkels from the group of Technology and sustainability studies at Eindhoven University of Technology, respectively

[34] This attribute was constructed to replace 'Euro savings' and 'continuous costs' after the preliminary experiment (see Appendix B)

Utility model calibration

Appendix C

The recommender system employed in the experiments of this thesis uses Multi-Attribute Utility Theory as its decision-making strategy. In MAUT, user-assigned weights are multiplied with attribute values of an option, and then summed to get the utility of this option for this user. However, the attributes that are specified for the energy-saving measures do not have the same scale (see Table 26 in Appendix B).

Incompatible scales are usually not a problem in MAUT because they are normalized by user weights: wider scales just have lower weights. In the recommender system employed in this research, however, users can increase their weights using clicks. In this case, a click to increase the weight of 'Effort once' should ideally have a similar effect as a click for 'Euro savings', similar, that is, to the perception of the user. The attribute values therefore need to be normalized before they can be subjected to MAUT.

Since the trade-off between these heterogeneously scaled attributes is a subjective issue, an experiment was devised to find the relative weight of each attribute. Note that the procedure of this experiment closely follows the experiment conducted by Poortinga et al. (2003).

Procedure

Task

Participants were asked to rate 33 *fictitious* energy-saving measures on a 1 (very unattractive) to 8 (very attractive) scale (other values on the scale had a number but no qualitative label). The measures were composed of the attributes listed in Table 27. In order to avoid familiarity effects, the measures did not include a name or description.

The 33 fictitious energy-saving measures followed a 7-variable, 4-level fractional factorial design. 32 of the 33 fictitious measures were 'manipulations', the last one had mean values on all attributes (a center point), and was presented five times at equal intervals. In effect, each participant completed 37 rating tasks.

The measures had fictitious values on all attributes that followed an orthogonal design. In an orthogonal design, attribute levels are chosen in a way that optimizes the power of finding a main effect in a regression analysis. Because there was a dependency in the attributes in the form of *time before return of investment = 12 * cost once / (euro savings – continuous costs)*, the attribute 'continuous costs' was made contingent on the other attributes. Also, in some energy-saving measures, an attribute value was replaced by the center-point value in order to remove implausible measures.[35]

Before the experiment, participants were shown an instruction and a real (named) energy-saving measure as an example, c.f. a programmable thermostat, with values Effort once = 19,

[35] C.f. measures with negative continuous costs.

continuous effort = 6, cost once = 48 Euros, continuous costs = 0 Euros/year, euro savings = 58 Euros/year, kilowatt-hour savings = 644 kWh/year, comfort= 10, and environmental effects = 3.

Participants

23 participants were recruited from the pool of friends and extended family of the researchers. Attention was paid to get a good mixture of sexes (11 female, 12 male), ages (M = 31.8, SD = 14.9) and occupations (14 students, 7 employed, 2 retired). All participants participated on a voluntary basis.

Measures

Besides the demographics reported above, the only measure taken was a 1 to 8 score on each of the 37 rating tasks. These scores could be subjected to a linear mixed regression analysis, the B weights of which could be used for normalizing the attribute values of the real energy-saving measures used in the next experiment.

Hypotheses

Direction of effect

Although the experiment was of an exploratory nature, some predictions could be made with respect to the direction of the effect on the assigned score. Specifically, effort, cost and time before return of investment were expected to have a negative effect, while savings, comfort and environmental effects were expected to be positively related to the assigned score.

Results

First analyses – reinterpretation

In a first analysis, a linear mixed regression was performed to predict the assigned score using the variables 'effort once', 'continuous effort', 'cost once', 'Euro savings', 'kilowatt-hour savings', 'time before return of investment', 'comfort' and 'environmental effects'[36]. 'Effort once' and 'Euro savings' were found to have an insignificant effect on assigned score, which means that these attributes were effectively of no importance to the participants. We predicted that users internally extracted the 'continuous costs' from 'Euro savings' to get the actual yearly Euro savings.

In a second analysis, we therefore replaced 'Euro savings' by a constructed variable of 'Euro savings' minus 'continuous costs', which we named 'real Euro savings'[37]. This constructed variable produced a significant effect. This means that participants calculated the actual savings that a measure would provide, and that they used this constructed value to evaluate

[36] The attribute 'continuous costs' was not included in the analysis in order to prevent multi-collinearity.

[37] In this analysis, the attribute 'cost once' was excluded to prevent multi-collinearity.

the measure's savings. We decided that the final experiment should include this construct, and not the original 'Euro savings' and 'continuous costs'.

In the second analysis 'effort once' was still not significant. Most likely, this was the result of the fact that our participants paid a lot more interest in continuous effort, to the effect that this attribute ultimately overshadowed the 'effort once' attribute.

Final analysis – normalization values

The B weights of the final analysis, reported in Table 27, were used to normalize the attribute values of the real energy-saving measures. Since the final analysis could not include the 'cost once' variable because of multi-collinearity, the normalization weight of this attribute was calculated to be $B_{real\ Euro\ savings} * (B_{time\ before\ return\ of\ investment} / 12)$.

Table 27: Normalization weights for the attributes of our energy saving measures

Variable	Scale	Manipulated range	B weight	Normalization weights
Effort once	0 to 50	0, 15, 25, 35, 50	-.00281	-.00281
Continuous effort	0 to 50	0, 15, 25, 35, 50	-.0222	-.0222
Cost once	Euros	5, 7.5, 15, 25, 75	calculated	-.000105
Continuous costs	Euros / year	Contingent	n/a	n/a
Euro savings	Euros / year	10, 25, 50, 100, 150	n/a	n/a
Kilowatt-hour savings	kWh / year	10, 40, 200, 400, 2000	.000130	.000130
Time before return of investment	months	3, 6, 12, 24, 36	-.0197	-.0197
Comfort	-25 to 25	-25, -10, 0, 10, 25	.0270	.0270
Environmental effects	-25 to 25	-25, -10, 0, 10, 25	.0377	.0377
'real' Euro savings	Euros / year	n/a	.00638	.00638

Evolution of the system – design and user tests

Appendix D

The Web Recommender System developed for the experiments of this thesis were to be used in 'unsupervised' online experiments. Therefore, adequate user interface design was critical, as participants were not able to ask the experimenter questions.

The system underwent an elaborate series of design improvements that were meant to increase usability on the one hand, and to ensure a stable experimental procedure on the other hand. Several 'snapshots' of the development process will be displayed and discussed in this appendix.

First 'working' version

The first working version (filled with debug data) showed three tables with choice options, divided between the 'recommendations', the 'trade-offs' and the 'other options'. The options could be selected and 'put in shopping cart', or marked as 'already applied'. Chosen measures were shown in the table with the heading 'chosen measures'.

Preference elicitation only allowed the *increase* of preference weights. In the attribute preference elicitation, this could be done by clicking on the attributes displayed in the table with the heading 'your preference'. In the case-based preference elicitation, this could be done by clicking on 'more like this' next to the 'trade-offs' table.

The main problem with this design was the use of ambiguous labeling: words like 'recommendations' and 'trade-offs' are recommender system jargon, and 'shopping cart' reminded of a web shop environment.

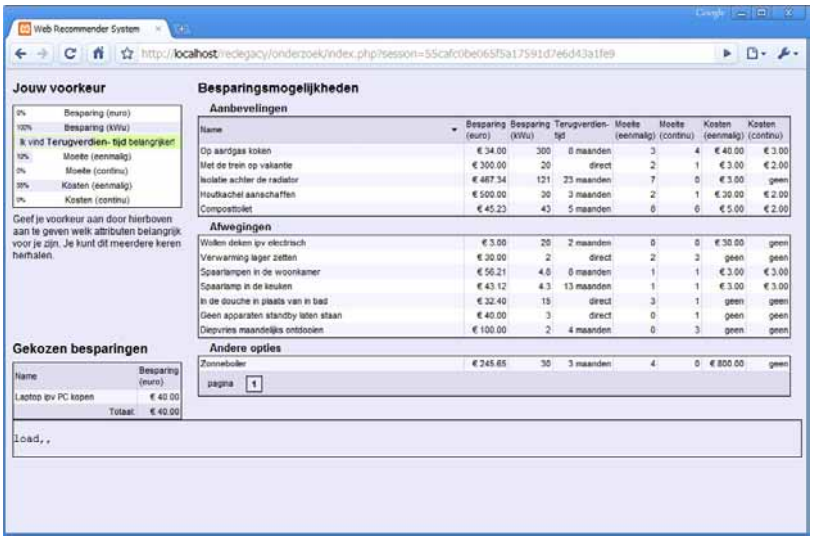


Figure 19: First working version, attribute-based preference elicitation

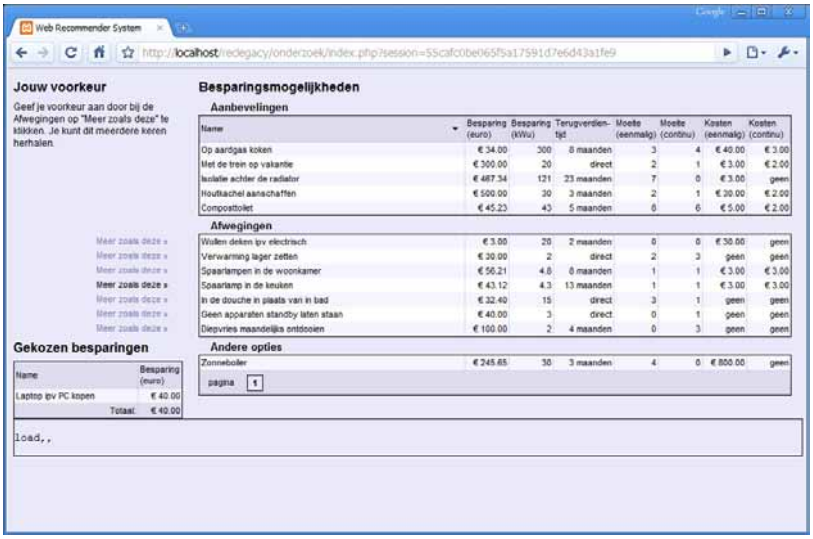


Figure 20: First working version, case-based preference elicitation

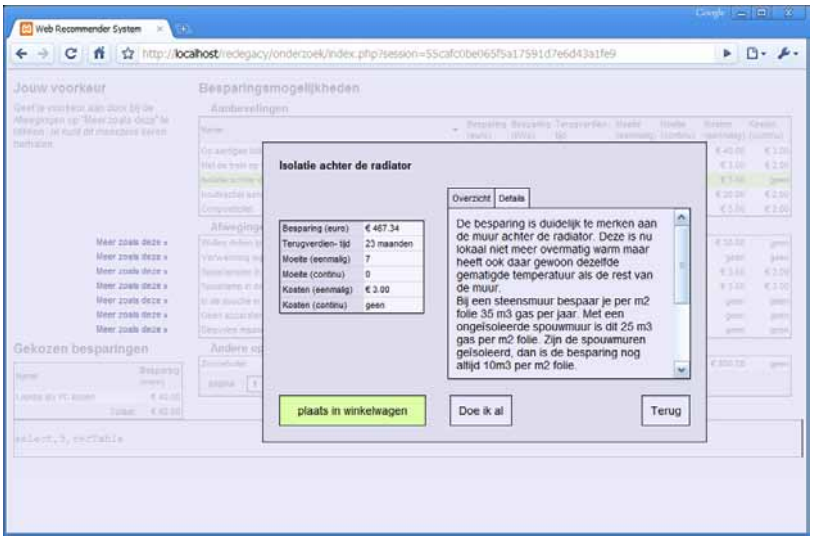


Figure 21: First working version, selected item

Version used in pre-test 1

Before the first pre-test with real users, the interface was changed on several points (besides loading the 'real' energy-saving measures):

- Each section was indicated more clearly with an icon
- Values of 'subjective' attributes were displayed graphically with linear scales
- Labels were changed: 'recommendations' became 'this fits you', 'trade-offs' became 'you can also consider this', 'your preference' became 'this is important to me', 'more like this' became 'this fits me' and 'put in shopping cart' became 'I want to do this'
- Measures that were 'already applied' (now labeled 'I'm already doing this') were now also displayed in the table with chosen savings (with a separate sub-total)

The system was tested with 4 users, using a think-aloud protocol to capture any misconceptions that users could have about the workings of the system. The biggest problem for the experiment was the fact that the test users only sporadically used the preference elicitation facilities, but instead systematically evaluated all available choice options. We concluded that a 'commercial' implementation of this system would have at least a tenfold of the number of choice options, which would render the systematic approach unfeasible. We therefore sought to prevent this behavior in the next version.

Furthermore, users complained about the fact that they could only increase their preference, and not decrease it. Moreover the workings of the case-based preference elicitation was not very clear, specifically, not all users understood the difference between selecting an item from the second table and indicating 'this fits me'.

Finally, users did not notice the button that could be used to end the experiment in the top right corner of the interface.

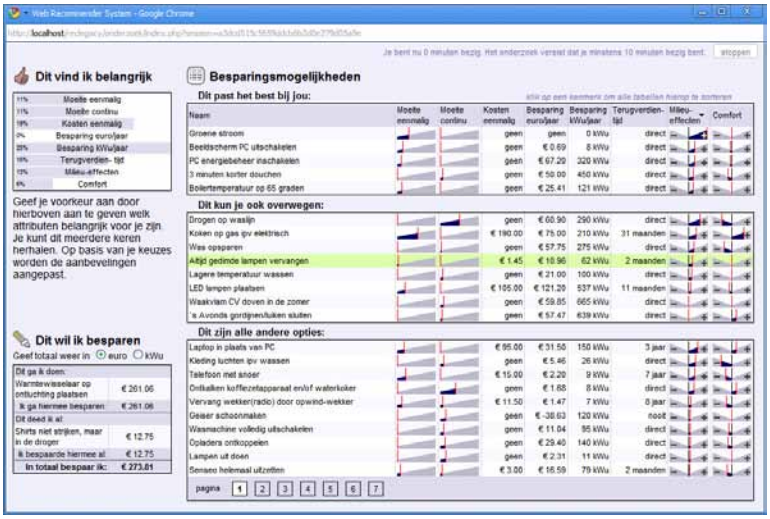


Figure 22: Version used in pre-test 1, attribute-based preference elicitation

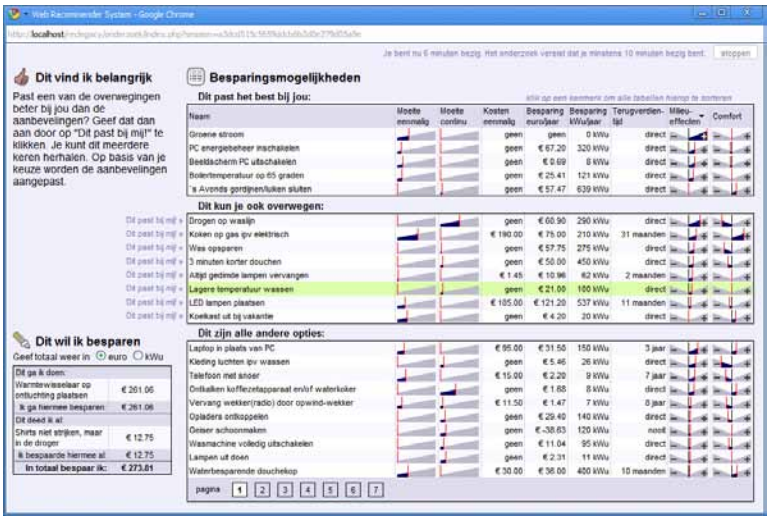


Figure 23: Version used in pre-test 1, case-based preference elicitation

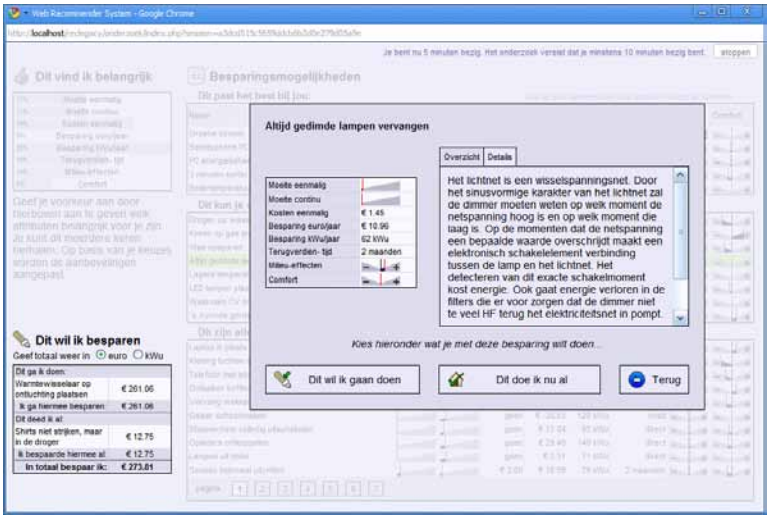


Figure 24: Version used in pre-test 1, selected item

Version after pre-test 1, used in pre-test 2

Based on the results of pre-test 1, the interface was significantly improved on the following points:

- We removed the 'other options' table; only a subset of the choice options was shown to the user, who was now 'forced' to use the preference elicitation more extensively
- We allowed users of both preference elicitation methods to decrease as well as increase their preference for a certain attribute or case
- Extra borders increased the visibility of the case-based preference elicitation buttons; the description more explicitly indicated the effect of these buttons
- The 'stop' button was moved from top-right to bottom-left, and got a more noticeable background color when active
- The tables with chosen and 'already applied' measures were placed in the spot that used to hold the 'other options' table, giving them more horizontal space

We again tested the system with 4 users, using the think-aloud protocol to signal problems. These tests found that the interaction had significantly improved, but that there were still some issues with the case-based preference elicitation and the second ('trade-offs') table. Specifically, the difference between the recommendations in the first and second table was not clear, and it was also still not clear how the case-based preference elicitation buttons related to the second table.

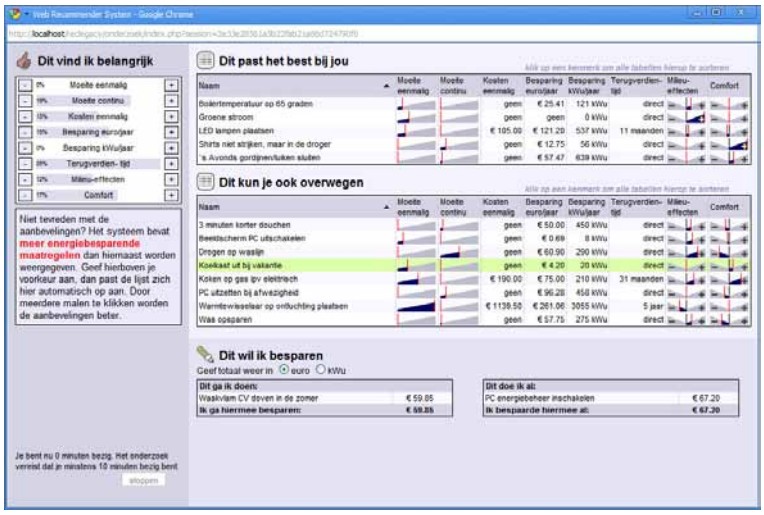


Figure 25: Version used in pre-test 2, attribute-based preference elicitation

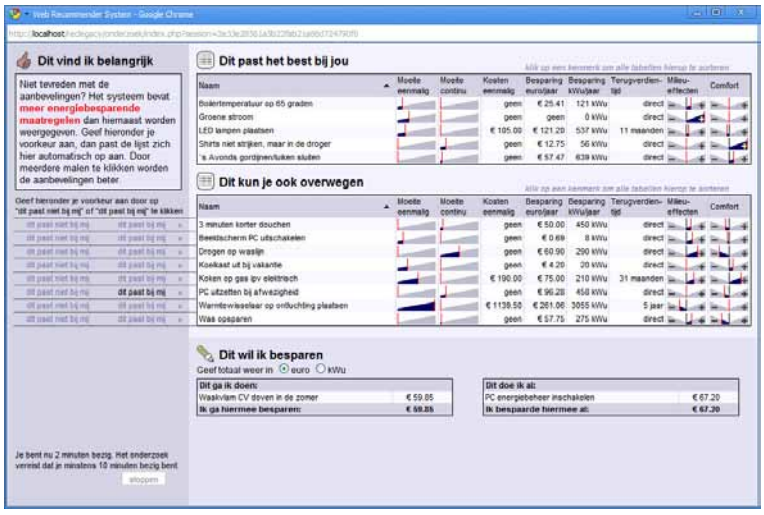


Figure 26: Version used in pre-test 2, case-based preference elicitation

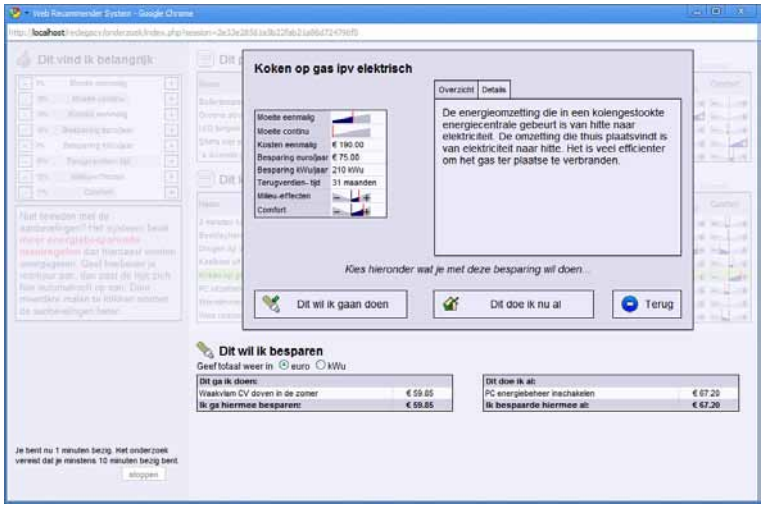


Figure 27: Version used in pre-test 2, selected item

Version after pre-test 2, used for experiment 1

Based on the second pre-test, we 'slimmed down' the interface even more, to make each component more understandable. Specifically:

- The interface was explicitly divided into three steps: preference elicitation, choice, and review; users were instructed to repeatedly follow these steps
- Increasing and decreasing preference weights was displayed more explicitly with red and green buttons showing 'thumbs down' and 'thumbs up'
- The 'trade-off' table became an integral part of the case-based preference elicitation method; we only showed it in this method, and the items were not clickable anymore, only the 'this fits me' and 'this does not fit me' buttons
- Consequently, users could only choose from the 'recommendations' table, which was moved to the center of the interface

This interface was tested with two users, and no problems occurred. The interface was therefore deployed online, and used in experiment 1.

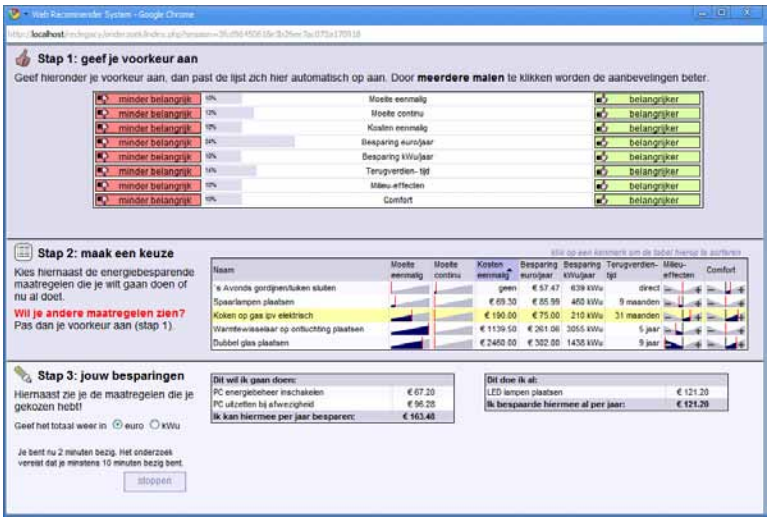


Figure 28: Version used in experiment 1, attribute-based preference elicitation

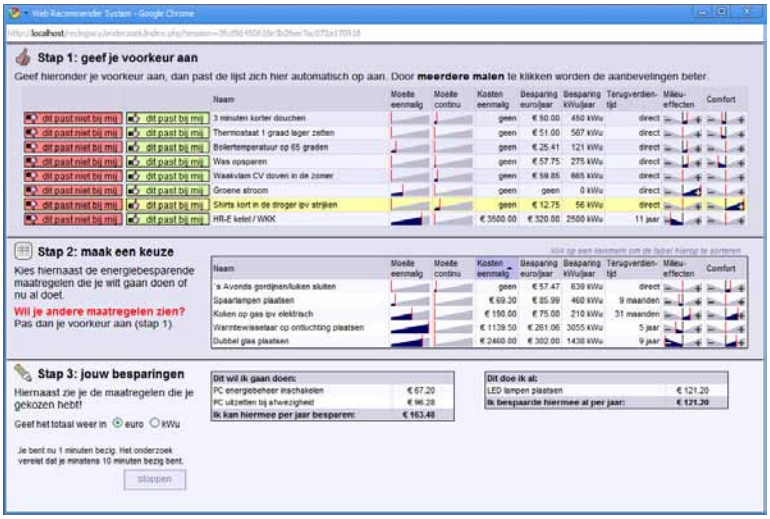


Figure 29: Version used in experiment 1, case-based preference elicitation

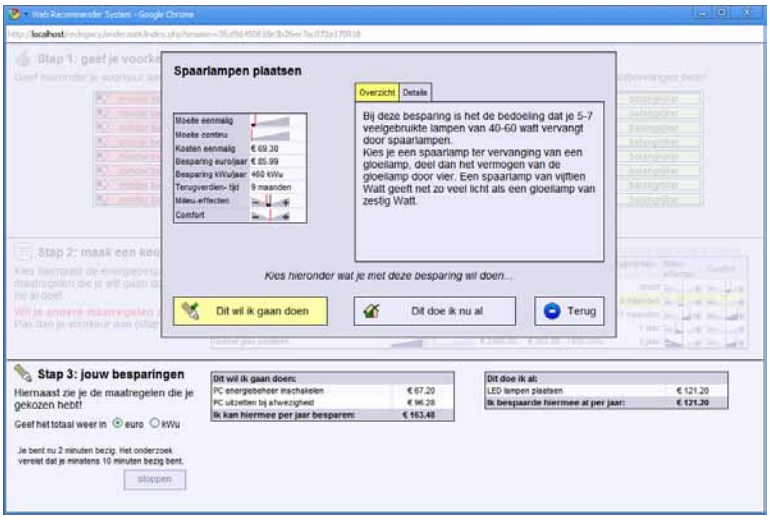


Figure 30: Version used in experiment 1, selected item

Version used in experiment 2

Based on feedback gathered in experiment 1 (through one open question in the post-experimental questionnaire, as well as replies on the various forums that were used to distribute the experiment), we slightly updated the interface for experiment 2, besides the addition of an agent. Specifically:

- We included 'live' help in the system at various points, so that users could review the instructions that were given before the interaction
- The attribute-based preference elicitation method included extra buttons for increasing/decreasing preference in larger steps, as some users complained about having to click too many times; we consciously avoided the use of sliders, so that click streams would still provide sufficient data about the use of the preference elicitation

Screenshots of the system used in experiment 2 can be found in Figure 1 and Figure 2.

The system – technology

Appendix E

The Web Recommender System was built as a generic adaptive online recommender system. The system presents an AJAX interface generated by a PHP back-end in conjunction with a MySQL database. The system configuration is completely database-driven, meaning that no changes have to be made to the source code in order to implement the system with a different set of attributes, choice options, user models and adaptive features. The system used in the experiments in this thesis is therefore actually *a specific instantiation* of the Web Recommender System.

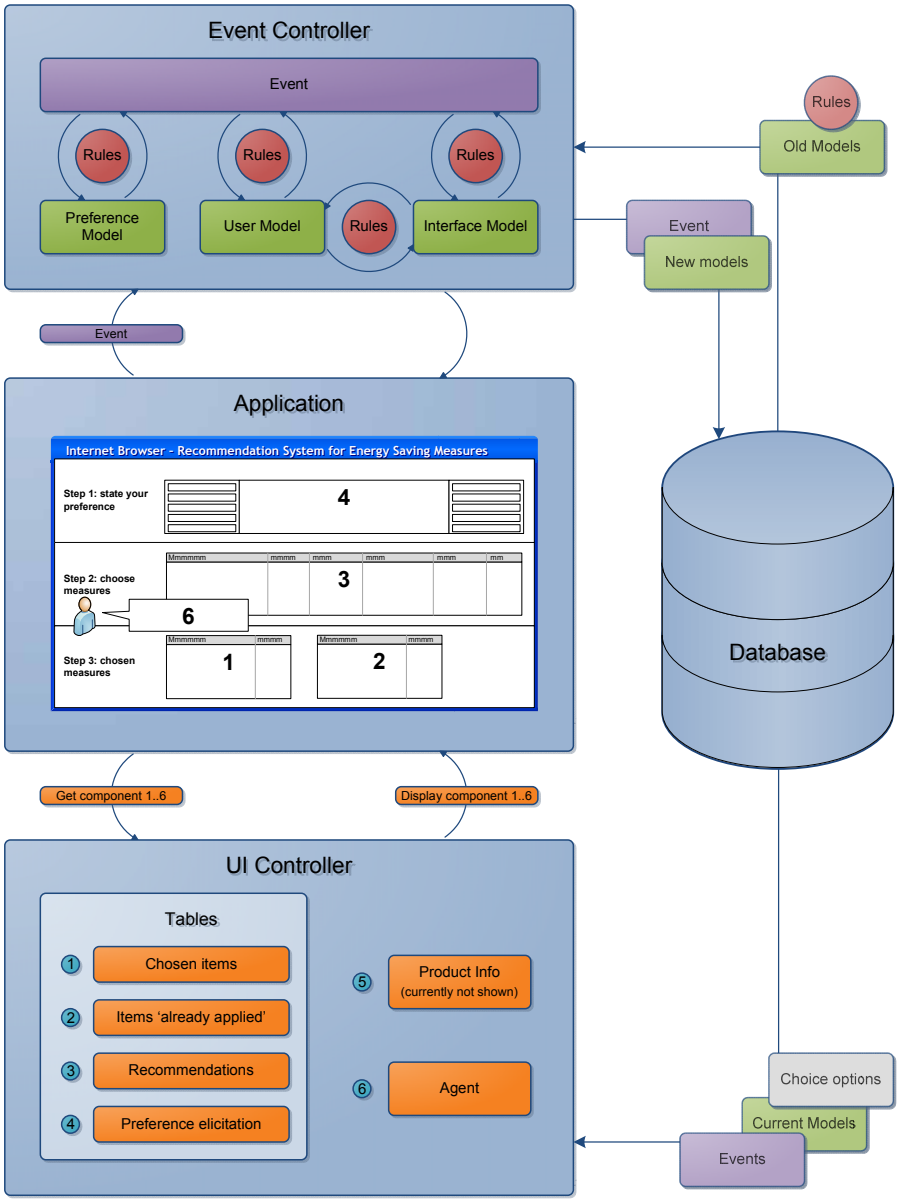


Figure 31: Schematical representation of the architecture of the Web Recommender System

Interaction cycle

Initialization

All interface updates in the Web Recommender System are database-driven. When a user boots up the system, the event controller creates a database record with the initial 'interface model' (a description of the current layout of the interface), 'user model' and 'preference model'[38]. Based on these models, the UI controller generates the initial interface and presents this to the user.

Event-cycle

When the user clicks somewhere in the interface, an event is generated and sent to the event controller. The event controller now first retrieves the old models and any applicable rules, and checks whether the rules warrant a change in the models. This works as follows:

- If the user clicks on a button in the preference elicitation, the preference model is updated.
- If a click matches a prediction rule (see Appendix G), the user model is updated (e.g.: a click that increases the preference weight for comfort also reduces the user model value for commitment).
- If this makes the user model pass a certain adaptation threshold, the interface model is updated accordingly (e.g. if commitment passes below -0.7, the total savings will be displayed in Euros).
- Finally, if the user explicitly changed an interface feature, the interface model is updated to match this change (e.g. if the user changed the display of information from general to details).

The event and the updated models are consequently saved as a new entry in the database, and the event controller now tells the application to update the interface.

UI-cycle

The application now requests an update for the six parts of the interface. Updates are performed asynchronously and without locking the system. Each component checks its current state based on the updated interface model retrieved from the database. The recommendations are computed performing MAUT-based calculations on the choice options and the current attribute weights.

The generated interface-parts are sent back to the application which inserts them into the interface using a Javascript procedure that prevents the usual 'blank screen' that is usually experienced on a page refresh. The system is now ready, and the user can click another button, starting the event-cycle again.

[38] Experimental conditions are created by separate initial models for each condition

Advantages and drawbacks of the current system architecture

The main advantage of the current system architecture is the fact that adaptation can be provided in a way that is both generic and manageable. All the rules that comprise the adaptive behavior are saved in a database table that is comprehensive and easy to change.

Furthermore, the stateless design is robust against connection errors and system crashes. Since the current state of the interface is generated completely from scratch based on a database entry, the system will always pick up where it left off when the user got disconnected.

Finally, since every event is saved in the database with its accompanying models, it is easy to mine the process data and even to simulate the effects of new adaptations by rerunning saved interaction patterns on the changed system.

The main drawback of the current architecture is the fact that it makes extensive use of database queries, which reduces the scalability of our solution. Every click in the interface generates between twenty and fifty database queries, and over the course of our two experiments, over two million database queries were performed by the system. To prevent any problems during our experiment, we made extensive use of custom indexes, manual query optimization and caching. This reduced the execution time of virtually all queries to less than five milliseconds. Hash tables and other database performance techniques could further increase the scalability of the system. However, we predict that a real system with, say, ten thousand users a day, would need a different architecture that has less scalability issues.

Pre- and post-experimental questionnaires

Appendix F

The pre- and post-experimental questionnaires are displayed in Table 28 through Table 35. All questions use a five-point scale, except for 'satisfaction with the system', which consists of a nine-point scale. The columns '1' and '2' indicate in which experiment the question was asked. Questions are grouped by construct. The pre-experiment questionnaires also have a column in which other research is cited that also uses the item; post-experimental questionnaires (besides the QUIS) were developed for our experiments, and were therefore not based on specific related work.

Table 28: Pre-experimental questionnaire for 'commitment'

Dutch wording	Translation	Scale	1	2	Source
Het maakt niet uit hoeveel je bespaart, als je maar bespaart.	Save regardless of how much	disagree/agree	✓		
Het milieu redden is belangrijker dan een besparing in geld.	Savings more important than money	disagree/agree	✓		
Bijna alles wat mensen doen is slecht voor het milieu.	People are bad for environment	disagree/agree	✓		(Stern, 2000; Dunlap et al., 2000)
Lang niet alle energiebesparende maatregelen zijn de moeite van het uitvoeren waard.	Not all savings are worth the effort	disagree/agree	✓		(Van Raaij & Verhallen, 1983)
Economische ontwikkeling is schadelijk voor het milieu.	Economy is bad for environment	disagree/agree	✓		(Stern, 2000)
Mensen maken zich veel te druk over het milieu.	People worry too much about the environment	disagree/agree	✓	✓	(Stern, 2000)
Ik spoor anderen aan om energie te besparen.	I encourage other to save energy	disagree/agree	✓		(Kaiser, 1998)
Ik ben dagelijks met energiebesparing bezig.	I'm saving energy every day	disagree/agree	✓		
Ik stoor me aan de overdreven aandacht voor energiebesparing.	Energy savings gets too much attention	disagree/agree	✓	✓	(Dunlap et al., 2000)
Als een bepaalde energiebesparende maatregel veel inspanning kost om in te zetten, dan vind ik dat:	When savings cost effort this is:	annoying/ not annoying	✓	✓	
Als een bepaalde energiebesparende maatregel veel geld kost om in te zetten, dan vind ik dat:	When savings cost money this is:	annoying/ not annoying	✓	✓	(Stern, 2000; Barr et al., 2005)
Als een energiebesparende maatregel mijn comfort vermindert, dan vind ik dat:	When savings reduce comfort this is:	annoying/ not annoying	✓		(Stern, 2000; Van Raaij & Verhallen, 1983; Barr et al., 2005)
Als ik meer belasting moet betalen voor een beter milieu, dan vind ik dat:	Paying more taxes for environment is:	annoying/ not annoying	✓		(Stern, 2000)

Table 29: Pre-experimental questionnaire for 'domain knowledge'

Dutch wording	Translation	Scale	1	2	Source
Ik let er continu op hoeveel energie ik verbruik.	I always pay attention to my energy usage	disagree/agree	✓		
Ik weet precies hoeveel energie elk apparaat in mijn huishouden verbruikt.	I know energy consumption of all devices	disagree/agree	✓	✓	
Ik begrijp het onderscheid tussen verschillende soorten energiebesparende maatregelen.	I understand difference between measures	disagree/agree	✓	✓	
Ik kan de voor- en nadelen van een gegeven energiebesparende maatregelen afleiden na het lezen van een korte beschrijving.	I can understand pros and cons of measures	disagree/agree	✓		
Alle verschillende manieren van energiebesparing komen uiteindelijk toch op hetzelfde neer.	All measures are eventually the same	disagree/agree	✓		
Ik ben bekend met energiebesparende maatregelen waar de meeste mensen nooit van gehoord hebben.	I know more measures than others	disagree/agree	✓	✓	(Flynn & Goldsmith, 1999)
Ik zoek vaak naar extra informatie over interessante energiebesparende maatregelen.	I search for extra info about measures	disagree/agree	✓		
Ik begrijp niets van de meeste energiebesparende maatregelen.	I don't understand most measures	disagree/agree	✓		(Flynn & Goldsmith, 1999)
Ik weet welke energiebesparingen zinvol zijn om uit te voeren.	I know which measures are useful	disagree/agree	✓	✓	(Flynn & Goldsmith, 1999)
De term "ecologische voetafdruk" is voor mij:	Term "ecological footprint" is:	unfamiliar/ familiar	✓		
De term "koolstofkringloop" is voor mij:	Term "carbon cycle" is:	unfamiliar/ familiar	✓		
De term "sluipverbruik" is voor mij:	Term "energy leakage" is:	unfamiliar/ familiar	✓		
Bij het kiezen van energiebesparende maatregelen vertrouw ik op mijn eerste gevoel.	I trust my instincts when choosing measures	disagree/agree	✓		
Ik ben in staat om verschillende energiebesparende maatregelen tegen elkaar af te wegen.	I can make trade-offs between measures	disagree/agree	✓		(Flynn & Goldsmith, 1999)
Als ik een energiebesparende maatregel ga uitvoeren, is dit een afgewogen keuze.	When I implement a measure, it's a conscious trade-off	disagree/agree	✓		
Ik ben in staat om goede energiebesparende maatregelen te selecteren.	I can choose the right measures	disagree/agree	✓		(Flynn & Goldsmith, 1999)
Ik twijfel wel eens of ik goede energiebesparende maatregelen heb gekozen.	I doubt whether I choose the right measures	disagree/agree	✓		
Ik denk dat er betere energiebesparende maatregelen bestaan dan de maatregelen die ik zelf doe.	I think there are better measures	disagree/agree	✓		

Table 30: Post-experimental questionnaire for 'satisfaction with the system'[39]

Dutch wording	Translation	Scale	1	2
Het systeem is:	The system is:	terrible/wonderful	✓	✓
Het systeem is:	The system is:	complex/easy	✓	✓
Het systeem is:	The system is:	frustrating/ satisfying	✓	✓
Het systeem is:	The system is:	dull/stimulating	✓	✓
Het systeem is:	The system is:	rigid/flexible	✓	✓

Table 31: Post-experimental questionnaire for 'satisfaction with the chosen measures'

Dutch wording	Translation	Scale	1	2
Ik ben blij met de maatregelen die ik gekozen heb.	I like the measures I've chosen	disagree/agree	✓	✓
Ik denk dat ik de beste maatregelen uit de lijst heb gekozen.	I think I chose the best measures	disagree/agree	✓	✓
De door mij gekozen maatregelen passen precies bij mij.	The chosen measures fit my preference	disagree/agree	✓	✓
Hoeveel van de door jou gekozen maatregelen ga je daadwerkelijk uitvoeren?	How many measures will you implement	none/all	✓	✓

Table 32: Post-experimental questionnaire for 'perceived usefulness'

Dutch wording	Translation	Scale	1	2
Het systeem heeft mij milieubewuster gemaakt.	The system made me more energy-conscious	disagree/agree	✓	✓
Het systeem beperkte me in mijn vrijheid om keuzes te maken.	The system restricted my choice freedom	disagree/agree	✓	✓
Ik zou dit systeem vaker gebruiken als dat mogelijk was.	I would use the system more often	disagree/agree	✓	✓
Met dit systeem kan ik beter milieuvriendelijke keuzes maken.	I make better choices with the system	disagree/agree	✓	✓
Ik vond het systeem nutteloos.	The system was useless	disagree/agree	✓	✓
Ik zou dit systeem aan anderen aanraden.	I would recommend the system to others	disagree/agree	✓	✓
Het systeem begreep mijn voorkeur volledig.	The system understood my preference	disagree/agree	✓	✓
Het systeem gaf slechte aanbevelingen.	The system made bad recommendations	disagree/agree	✓	✓
De aanbevelingen van het systeem pasten bij mijn voorkeur.	The recommendations fitted my preference	disagree/agree	✓	✓

[39] Based on the QUIS, this questionnaire can be found at http://hcibib.org/perlman/question.cgi?form=QUIS.

Table 33: Post-experimental questionnaire for 'understandability'

Dutch wording	Translation	Scale	1	2
Ik vond het systeem eenvoudig in het gebruik.	The system was easy to use	disagree/agree	✓	✓
Ik raakte in de war door dit systeem.	The system confused me	disagree/agree	✓	✓
Ik begreep niets van dit systeem.	I didn't understand the system at all	disagree/agree	✓	✓
Ik begreep goed hoe ik mijn voorkeur kon aangeven.	I understood how to indicate my preference	disagree/agree	✓	✓
Hoe moeilijk of makkelijk vond je het om met hulp van dit systeem energiebesparende maatregelen te vergelijken?	How difficult/easy was comparing measures?	difficult/easy	✓	✓
Hoe moeilijk of makkelijk vond je het om je voorkeur aan te geven in het systeem?	How difficult/easy was stating preference?	difficult/easy	✓	✓
Hoe moeilijk of makkelijk vond je het om verschillende attributen van de energiebesparende maatregelen te vergelijken?	How difficult/easy was comparing attributes?	difficult/easy	✓	
Ik heb vooral naar de naam van de maatregelen gekeken, en nauwelijks naar de overige attributen	I looked primarily at name of measures, not attributes	disagree/agree	✓	

Table 34: Post-experimental questionnaire for 'perceived personal help'

Dutch wording	Translation	Scale	1	2
Het systeem is:	The system is:	dumb/smart		✓
Het systeem is:	The system is:	not helpful/helpful		✓
Het systeem is:	The system is:	trustworthy/ not trustworthy		✓
Het systeem denkt met mij mee.	The system thinks my way	disagree/agree		✓
Het systeem doet wat ik wil.	The system does what I want	disagree/agree		✓
Het systeem past zich aan mij aan.	The system adapts to me	disagree/agree		✓
Het systeem en ik vormden een team.	The system and I were a team	disagree/agree		✓
Het systeem bood persoonlijke hulp.	The system gave personal help	disagree/agree		✓

Table 35: Post-experimental questionnaire for 'acceptance and understanding of adaptive behavior'

Dutch wording	Translation	Scale	1	2
Ik begrijp hoe het systeem werkt.	I understand the system	disagree/agree		✓
Ik begrijp waarom aanpassingen gedaan werden.	I understand why adaptations were made	disagree/agree		✓
De aanpassingen die het systeem maakte waren:	The adaptations were:	unexpected/natural		✓
De aanpassingen die het systeem maakte waren:	The adaptations were:	unclear/clear		✓
De aanpassingen die het systeem maakte waren:	The adaptations were:	annoying/ not annoying		✓
De aanpassingen die het systeem maakte waren ongepast.	The adaptations were uncalled for	disagree/agree		✓
De aanpassingen die het systeem maakte hielpen mij.	The adaptations helped me	disagree/agree		✓

Making the system adaptive

Appendix G

This Appendix demonstrates how the process-rule values and adaptation threshold values were determined using the data gathered in experiment 1. A thorough description of the adaptive behavior can be found in the section 'Making the system adaptive' on page 37.

Process-rule values

Based on process data predictors of domain knowledge and commitment derived in experiment 1, we were able to define several process-rules that update the user model based on clicks in the interface. Besides process-rules derived in experiment 1, we also included 'corrective behaviors' in our user model; these are actions where the user manually changes something that could also be changed adaptively by the system. Such a change can be seen as a (pro-active) correction of a certain adaptation. For example, if the user changes the information display from detailed info to general info, this action should reduce the value of domain knowledge, since changing the information to general info is an adaptation for novice users (see paragraph 'Possible adaptations' below).

In order to find the optimal update values for each of these rules, we simulated the construction of user models on our process data of experiment 1. Specifically, we virtually 're-ran' the interaction of the participants of experiment 1, this time also constructing a user model for each of them. By correlating the constructed user model value for commitment and domain knowledge with the values obtained from the pre-experimental questionnaires, we were able to tweak the user model so as to find the optimal prediction-rule values. The rules and values found in this process are displayed in Table 36 for domain knowledge and Table 37 for commitment.

Table 36: Prediction rules for domain knowledge

User action[40]	Update value	Evidence
Increase weight of an attribute	+0.03 * units increase[41]	Table 8
Indicate 'already doing this'	+0.10	Table 8
Choose item	-0.05	Table 8
Indicate 'already doing this'	KWh-savings of the item / 10000	Table 10
Indicate 'already doing this'	(Cont. effort of the item - 25) / 250	Table 10
Indicate 'already doing this'	Comfort of the item / -400	Table 10
Change information from general info to detailed info	+0.25	Corrective behavior
Change preference elicitation method from attribute-based to case-based	-0.15	Corrective behavior

[40] Note that the inverse of each action causes an update with opposite value.

[41] As explained in the chapter 'An adaptive recommender system', the various buttons for increasing/decreasing attribute weights increased/decreased the weights by different amounts (1 unit, 2 units, or 5 units).

Table 37: Prediction rules for commitment

User action[42]	Update value	Evidence
Choose item	+0.05	Table 9
Choose item with KWh-savings › 2000	+0.20	Table 11
Choose item with environmental effects › 6	+0.10	Table 11
Indicate 'already doing this'	Environmental effects of the item / 250	Table 11
Indicate 'already doing this'	One-time cost of the item / 100000	Table 11
Indicate 'already doing this' for item with comfort › 10	- 0.05	Table 11
Increase weight of environmental effects	+0.10 * units increase	Table 12
Increase weight of continuous effort	-0.05 * units increase	Table 12
Increase weight of comfort	-0.05 * units increase	Model fitting
Change display of total savings from Euros to KWh	+0.25	Corrective behavior
Sort items by environmental effects	+0.25	Corrective behavior
Sort items by comfort	-0.25	Corrective behavior

Thresholds for adaptations

The threshold levels for the adaptations were determined in the simulation re-run of experiment 1 that was also used to determine the update values for the process-rules. The final threshold values as presented in Table 38 and Table 39 were selected in such a way that most users would eventually experience an interface fitted to their level of domain knowledge and commitment, taking care not to generate too many adaptations.

In order to prevent the interface from 'flipping' constantly when the user model fluctuates around a threshold, we put *hysteresis* in the threshold values, meaning that opposite adaptations would have different thresholds. For example, the interface switches from case-based preference elicitation to attribute-based preference elicitation when domain knowledge is higher than +0.70, but it switches back only when domain knowledge is lower than -0.70. It therefore rarely occurs that a certain feature flips back and forth within a short period of time.

Table 38: Adaptation thresholds for domain knowledge

Threshold	Adaptation
-0.70	Set PE method to case-based
-0.50	Set information type to general info
+0.50	Set information type to detailed info
+0.70	Set PE method to attribute based

[42] Note that the inverse of each action causes an update with opposite value.

Table 39: Adaptation thresholds for commitment

Threshold	Adaptation
-0.70	Display total savings in Euros
-0.50	Sort recommendations by comfort
+0.50	Sort recommendations by environmental effects
+0.70	Display total savings in KWh