Part IV

Final

10
Evaluating our Approach with a User Study

We have presented an approach for automated decision making, which involves contributions in three different directions: (i) representation of high-level qualitative preferences; (ii) preference-based decision making with user-centric principles; and (iii) explanations to justify choices. Our preference metamodel and language (Chapter 3) are justified by our study of how humans express preferences (Chapter 2), the decision making technique (Chapter 6) that is able to process such language was evaluated by a comparison with a human domain expert, and our explanation generation technique (Chapter 9) is justified by our investigation of how people explain their choices (Chapter 8). In this chapter, we evaluate through a user study these three parts of our work, focusing on the explanation and its impact on the trust and confidence of users on choices made by the decision making technique, and the comparison with existing explanation approaches. The description of our empirical evaluation is presented in Section 10.1, and its results are detailed in Section 10.2. Finally, we discuss the threats to the validity of our study in Section 10.3.

10.1
Study Description

This last study that we conducted in the context of this thesis consists of an experimental evaluation to which we adopted the same framework (Basili et al. 1986) used in our previous studies, detailed in Chapters 2 and 8. Our study goal, according the GQM template (Basili and Rombach 1988) is shown in Table 10.1.

The study we designed to achieve this goal consists of within-subjects user study, in which we use a developed application that allows participants to (i) express their preferences in a high-level language; (ii) receive a choice made by our decision making technique; and (iii) receive different explanations that justify this choice. With this application, we ask for participants' feedback with respect to these three dimensions and use this information to evaluate our approach. Moreover, the domain chosen for this study is *choosing a mobile phone to buy*, as mobile phones can be described in terms of attributes of different types, in general people have a

Definition element	Our experiment goal
Motivation	To assess the impact of different explanations on automated decision making,
Purpose	evaluate
Object	the user understanding, trust and confidence on choices made
Perspective	from a perspective of the researcher
Domain:people	as users receive explanations to justify those choices
Domain:system	from a decision making system
Scope	in a within-subjects study.

Table 10.1: Goal Definition (GQM template).

set of preferences they are aware of to make this choice (known preferences), and there are plenty of mobile phones available.

In next sections, we describe the research questions and hypotheses (Section 10.1.1), the procedure (Section 10.1.2) and participants (Section 10.1.3) of our study.

10.1.1 Research Questions and Hypotheses

Even though the main goal of this study is to investigate the impact of explanations on the user understanding, trust and confidence on choices made, and compare existing explanation approaches with respect to these dimensions, we also evaluate other aspects that our approach includes: (i) the language expressiveness; (ii) the decision making technique's choices; and (iii) the impact of explanations. Therefore, there are different research questions related to this study, presented below.

RQ1. Is our high-level preference language expressive enough for users to provide their preferences about a domain?

RQ2. Does our decision making technique make choices on user's behalf that they consider good?

RQ3. Do explanations increase the user understanding, trust and confidence on why a particular choice is made?

RQ4. Do different kinds of explanations (generated using using our approach (Chapter 9), Klein and Shortliffe's (Klein and Shortliffe 1994) and Labreuche's (Labreuche 2011)) have a different impact on the user understanding, trust and confidence on why a particular choice is made?

By answering these research questions, we are able to identify issues of our existing approach with respect to missing available preference types for users to

express themselves, and the quality of the decisions made by our technique, thus indicating points that should be addressed as future work. Moreover, we are able to conclude how much an explanation changes the evaluation of a choice by a user, and identify the best explanation generated by different approaches available.

But, as we mentioned before, our main focus is on the impact of explanations on automated decision making systems and the comparison between the different explanation approaches, and we list below our null hypotheses related to them.

$H1_0$: *The **choice quality** perceived by users on the choice made does not change after presenting them an explanation that justifies it.*

$H2_0$: *The **trust** of users in the choice made does not change after presenting them an explanation that justifies it.*

$H3_0$: *The user **decision confidence** on the choice made does not change after presenting them an explanation that justifies it.*

$H4_0$: *The three investigated kinds of explanations have the same impact on the **understanding of why (transparency)** why choices were made.*

$H5_0$: *The three investigated kinds of explanations have the same impact on the **choice quality** perceived by users on the choice made.*

$H6_0$: *The three investigated kinds of explanations have the same impact on the **trust** in choices made.*

$H7_0$: *The three investigated kinds of explanations have the same impact on the user **decision confidence** on choices made.*

10.1.2 Procedure

In order to answer our research questions and test our hypotheses, we have designed a user study in which participants interact with a developed software system, which implements the three components of our approach, namely the preference language, the decision technique and the explanation generation technique, and an interface to collect and display data. Moreover, it is also able to generate explanations with the approaches proposed by Klein and Shortliffe (Klein and Shortliffe 1994), and Labreuche (Labreuche 2011). The study consists of within-subjects comparing the impact of these three explanation approaches, besides analysing other aspects related to our decision making approach. Furthermore, our study involves making a decision about mobile phones, creating a hypothetical scenario in which participants are going to buy a mobile phone and need to choose one model from a set of available. As already explained, we chose mobile phones as the domain of the decision to be made as it fits the requirements of using our approach — most of people are aware of the attributes

that characterise mobile phones, and have preferences over individual attributes. Furthermore, we can retrieve a mobile phone database in a relatively easy way.[1] In addition, this domain is different from the domain of our previous studies — choosing a laptop (Chapter 2) and a hotel (Chapter 8) — from which we derived our preference language and explanation patterns, respectively.

Each participant has to go through seven steps while taking part of the study, each of which is described next. Screenshots of the application developed to be used in the study can be seen in Appendix D. The application has two available languages: English and Portuguese.

1. **Participant Data.** In order to collect demographic information of the participants, they are required to provide the following data: (i) age; (ii) gender; (iii) location (city and country); and (iv) working/studying field.

2. **Preferences.** The study participants are requested to imagine a situation in which they are going to buy a new mobile phone. In addition, in this scenario, they are provided with an intelligent system that will make a choice on their behalf and asks them to specify their preferences and restrictions over the mobile phone they want. Participants are able to specify their preferences using our language (presented in Chapter 6) through an interface that has many features, such as choosing explanation types with radio buttons, then selecting preferences parameters with combo boxes, setting preference formulae in a similar way to specifying rules in e-mail clients, and so on. Before providing their preferences, the participants receive a brief tutorial on how to interact with the interface and explanations about the language constructions. The application also is able to list the mobile phone attributes, and descriptions of each preference and priority type. For recording purposes, we store how long participants take to specify their preferences.

3. **Preference Language Evaluation.** After specifying their preferences, we request the participants to evaluate the interface and language they used in terms of two aspects: *perceived ease of use* and *perceived effort*, whose associated questions that are asked to participants are shown in the "Preference Language Evaluation" part of Table 10.2, and the possible answers for such questions are according to a 7-point Likert scale. These measured variables as well as others that are adopted in next steps of the study are part of a user evaluation framework of recommender systems (Chen and Pu 2010), whose questions were adapted to our study. Moreover,

[1]Imported from the Best Buy store (`http://www.bestbuy.com`), available through a REST API located at `https://bbyopen.com/developer`.

participants are requested to describe any preferences that they could not express in our language.

4. **Choice Analysis and Evaluation.** Based on the provided participants' preferences, we choose an option using our decision making technique, and present to the participant: (i) the chosen mobile phone; (ii) the next four mobile phones of the acceptable set ranked according to the decision function of our technique (so as to form five chosen options, which was deemed an adequate number in our previous study); and (iii) the remaining mobile phones initially hidden, but the participants can see them upon request to analyse all the 191 available mobile phones. Now, with this presented choice, participants are asked to evaluate it by answering the questions in the "Choice Evaluation" part of Table 10.2, which are related to variables: *choice quality*, *decision confidence* and *trust in choice*. They also have to specify which mobile phone they would choose, if they had to make the choice themselves.

5. **Explanation Impact.** After evaluating the chosen mobile phone, participants are provided with explanations to justify the choice made generated using our technique. The participants are then requested to answer questions in the "Explanation Impact and Explanation Comparison" part of Table 10.2, which are the same presented in the previous step together with a question related to the *transparency* variable. The goal of asking the same questions again is to evaluate if the choice quality perceived by users, and their trust and decision confidence change after receiving explanations. The questions related with these measures are initialised with the answers previously provided by the participants.

6. **Explanation Comparison.** The participants now receive the three possible explanations — generated by (i) our approach; (ii) Klein and Shortliffe's approach; (iii) Labreuche's approach — in a side-by-side comparison, and have to answer the same questions of the previous step. As our approach was already presented in the previous step, its answers are already initialised. Participants are requested to compare the three given explanations and evaluate them, and they are notified that they can review their opinion about the previously present explanation. In order not to create a pre-defined explanation order, we change the order of the explanations for different participants.

7. **Approach Evaluation.** Finally, participants have to answer final questions that evaluate the approach as a whole, which are shown in the "Approach Evaluation" part of Table 10.2.

Measured Variable	Question responded on a 7-point Likert scale from "strongly disagree" to "strongly agree"
Preference Language Evaluation	
Perceived ease of use	I find this interface to provide my preferences easy to use.
Perceived effort	Providing my preferences in this language required too much effort (*reverse scale*).
Choice Evaluation	
Choice quality	This application made really good choices.
Trust in choice	I feel that this application is trustworthy.
Decision confidence	I am confident that the choice made is really the best choice for me.
Explanation Impact and Comparison	
Transparency	I understand why the products were returned through the explanations in the application.
Choice quality	This application made really good choices.
Trust in choice	I feel that this application is trustworthy.
Decision confidence	I am confident that the choice made is really the best choice for me.
Approach Evaluation	
Perceived usefulness	This application is competent to help me effectively make choices I really like.
Intention to purchase	I would accept this choice if given the opportunity.
Intention to return	If I had to search for a product online in the future and an application like this was available, I would be very likely to use it.
Intention to save effort in next visit	If I had a chance to use this application again, I would likely make my choice more quickly.
Enjoyment	I found my visit to this application enjoyable.
Satisfaction	My overall satisfaction with the application is high.

Table 10.2: Measured Variables — adapted from (Chen and Pu 2010).

In the last four steps of the study, participants are also able to provide further comments. With this collected information we are able to extract issues of our current approach, evaluate our language, decision making technique, and existing explanation approaches (including ours). In next section, we present the participants involved in our study.

10.1.3 Participants

As in our previous studies, we selected participants using convenience sampling, by making invitations for volunteers via email to the social network of the researchers involved in this study. However, as participants were observed while taking part of the study, only participants in the same locations (in two

Gender		Male 21 (60%)		Female 14 (40%)	
City		Porto Alegre 19 (54.29%)	Rio de Janeiro 14 (40%)	Other 2 (5.71%)	
Age	<16 years 1 (2.86%)	16-25 years 9 (25.71%)	26-35 years 17 (48.57%)	36-45 years 2 (5.71%)	>45 years 6 (17.14%)
Field of Work of Study		Informatics 16 (45.71%)	Engineer 5 (14.29%)	Law 5 (14.29%)	Other 9 (25.71%)

Table 10.3: Demographic Characteristics of Participants.

different Brazilian cities) of the researchers could be selected — two participants were visiting from other locations. The demographic characteristics of the study participants, which are 35 in total, are described in Table 10.3.

10.2 Results and Analysis

Now, we will present the data collected in the study we performed, and discuss its results, which are split into four parts: (i) the analysis of provided preferences and the evaluation of the preferences language (Section 10.2.1); (ii) the evaluation of the choice made by our decision making technique and the impact of the given explanation (Section 10.2.2); (iii) the comparison between the different provided explanations (Section 10.2.3); and (iv) the evaluation of the overall approach (Section 10.2.4).

10.2.1 Preferences and Language Evaluation

After providing their personal data, participants had to specify their preference about mobile phones. As explained before, we briefly introduced to the participants the interface for providing preferences. The types of preferences (qualifying, rating, goal, etc.) were presented in a radio button list, from left to right (see Figure D.2(a) in Appendix D), and in order to avoid the bias of users always starting by the same first preference type (qualifying), we introduced the preferences in the opposite direction, starting from the *don't care*. On average participants took $15 min$ to specify their preferences, including the time to give the brief tutorial.

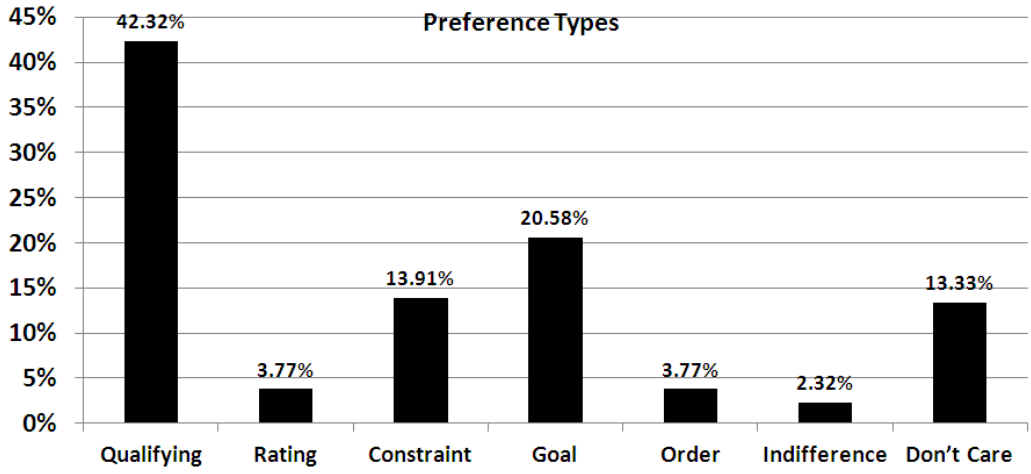
By observing the interaction of participants with the interface, we noticed they first took a few moments to get familiar with it, and then explored the available preference types, priorities and attributes. Many participants began providing their preferences by specifying the attributes they do not care about, for later concentrating on the characteristics they desire. As participants had the list of attributes available to them, they often looked at this list to remember their preferences — after providing a set of preferences, some checked the list again,

and said "*let me check if I forgot something*." People that have no or little interest in mobile phones seemed to not know what to specify, mainly because the attributes provided were not at a high level, such as "easy to use." Therefore, in order to explicitly provide preferences in this way, it is important for users to have at least some knowledge about the domain.

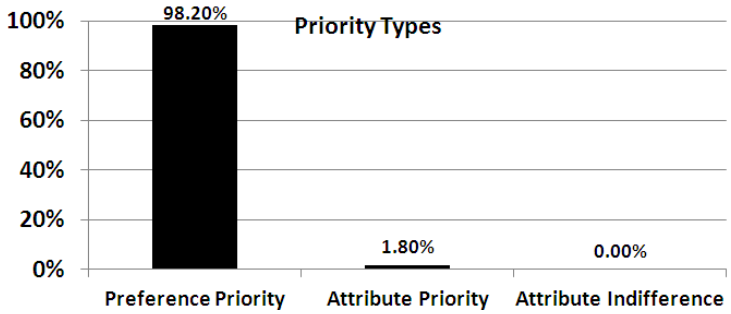
A set of histograms of the provided preferences is shown in Figure 10.1, where we present the distribution of the different preference types used, the different priority types used, and the expressive speech acts and rates used in qualifying and rating preferences. It can be seen in Figure 10.1(a) that *qualifying* is the most used preference type (42.32%), which indicate that people tend to use expressive-speech-act-based statements, which is one of the main advantages of our language with respect to existing preference languages. The most frequent set of preferences provided by participants is a combination of qualifying preferences with goals. The majority of participants — 26 (74.29%) — adopted preference priorities among preferences (i.e. numbered preferences), instead of priorities among attributes, and some split preferences into groups, that is, a set of preferences of priority 1, priority 2, and so on. Finally, the most common expressive speech act used was *want* (57%), but a wide variety of other expressive speech acts and rates were used.

Participants were then requested to evaluate their experience in providing preferences through the given interface and language, and the obtained results are shown in Figure 10.2. We observed during the execution of our study that it was not straightforward for older people (age > 45 years) to use the interface (including those who work with informatics), and also for them to provide their preferences as they are not familiar with the domain. Therefore, they are responsible for the worst scores with respect to the *perceived ease of use* (M = 5.63 and SD = 1.19) and the *perceived effort* (M = 3.06 and SD = 1.59) measurements. Even though the interface was considered ease to use for most of the participants, they reported that it is not intuitive, and without the given explanation it would not be easy to interact with it.

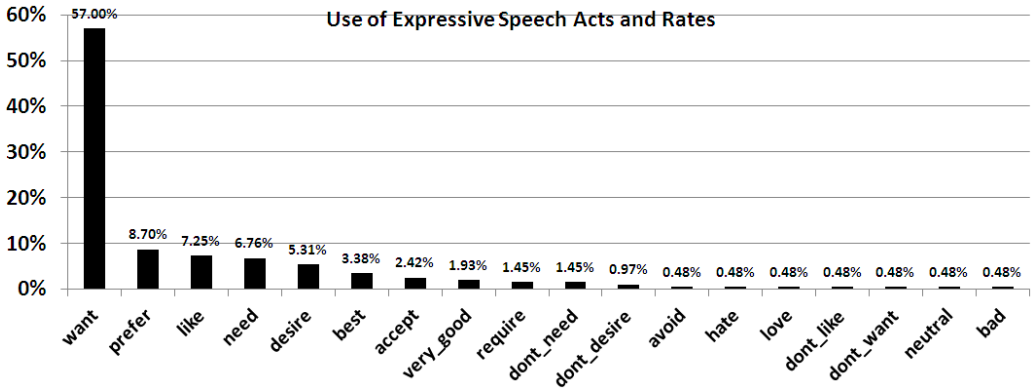
With regard to the effort spent in providing preferences, 45.72% participants (strongly) disagreed that providing preferences in our language requires too much effort. This issue can be divided into two parts: effort required to provide preferences and effort to express these preferences in our language. The additional question we asked of participants, which requests them to list any other preferences they wanted to express and they could not, helped us to distinguish what the participants meant. 25.71% of the participants pointed out limitations, but only 8.57% reported *language* limitations. Most of the limitations are related to missing attributes, such as alarm clock, operating system version and processor, or even subjective ones,



10.1(a): Preference Types.

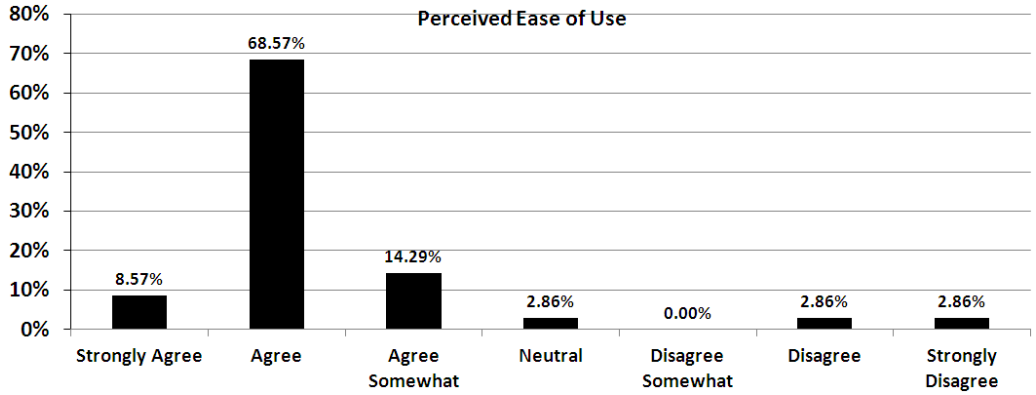


10.1(b): Priority Types.

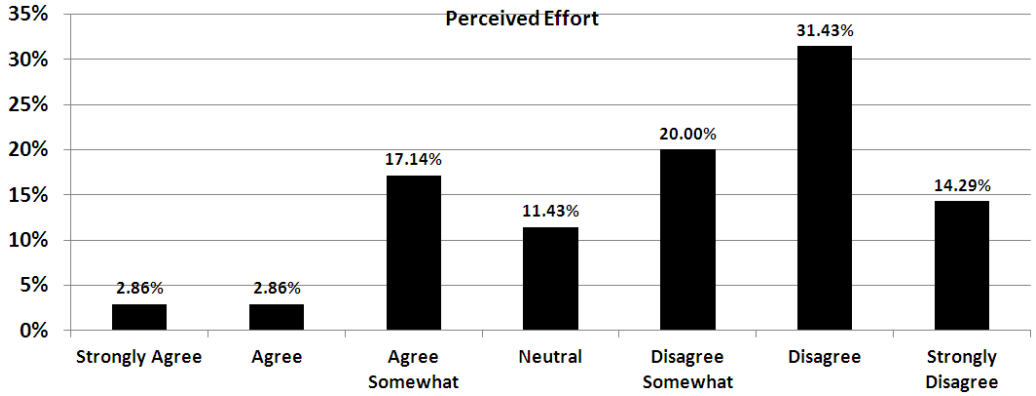


10.1(c): Use of Expressive Speech Acts and Rates.

Figure 10.1: Preference Analysis.



10.2(a): Perceived Ease of Use.



10.2(b): Perceived effort.

Figure 10.2: Language Evaluation.

e.g. usability, or ways of referring to attribute values, e.g. good or small, which are not missing preference constructions, but restrictions of the domain. Some of these restrictions are associated with elements that are part of our preference metamodel (e.g. scales and adjectives) that our decision making technique is unable to handle, and others are related to the engineering of the domain. Based on the given participants' comments, we make three observations.

1. It is important to make it explicit to users how they can express hard constraints. Some participants asked how they could express a restriction that *must* be satisfied, but they were not informed during the study. From these, some intuitively used *require* or *need*, which indicate hard constraints in our technique, but others did not.

2. It is interesting to consider the possibility of extending the language or having a user interface that provides "shortcuts" for the expression of more complete specifications to allow users to express a subset of the most important attributes, which was a limitation pointed out by one participant.

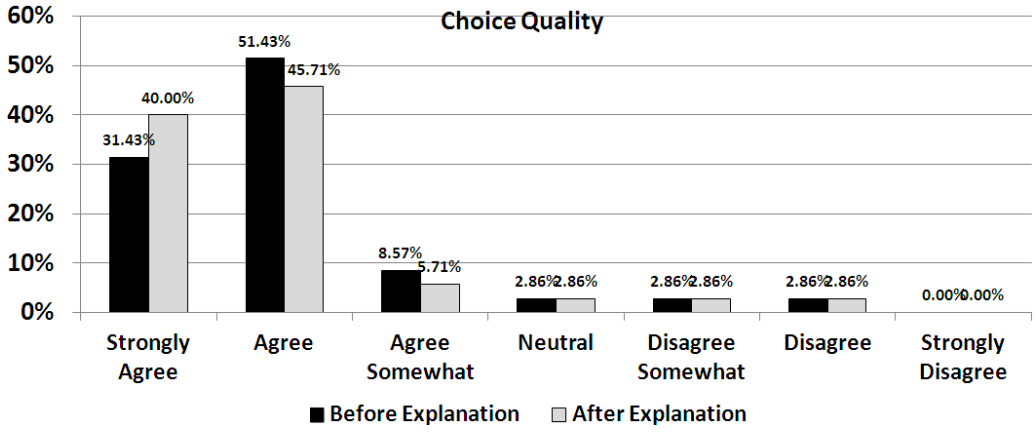
3. A single participant stated that the many different ways of expressing preference is confusing, while the others (97.14%) seemed to be comfortable with having many alternatives.

10.2.2 Choice Evaluation and Explanation Impact

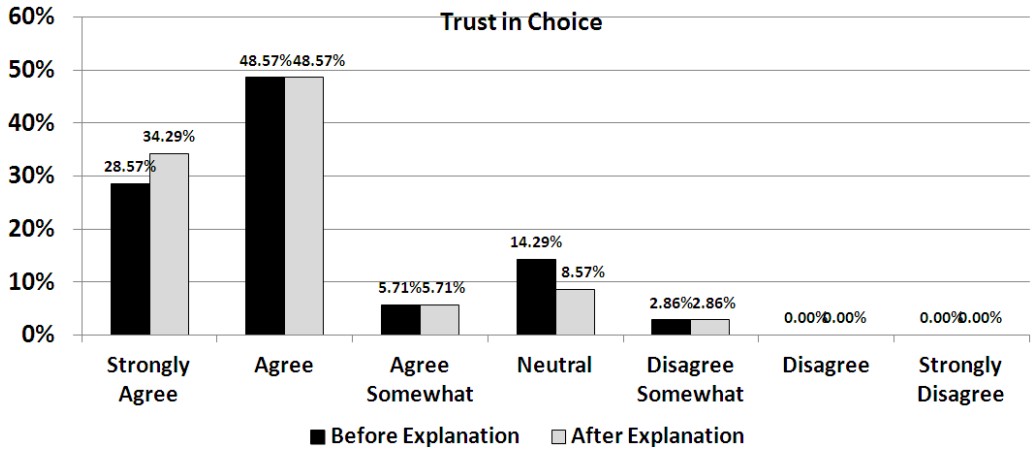
The next two steps of our experiment, namely choice evaluation and explanation impact, are discussed together. Based on the provided preferences, participants had a choice made on their behalf, and they had to evaluate this choice (and other selected options that are close to the chosen option) with respect to the *choice quality*, *trust in choice* and *decision confidence* — the obtained results are presented in Figure 10.3. It can be seen that our technique was evaluated with high levels of choice quality and trust, and also of decision confidence (but at a level lower than the other measurements) indicating that our technique is able to make adequate choices. We identified three situations in which the decision made was not good, detailed next.

- *Establishing a preference order without specifying preferred values.* If a preference like "I prefer X to Y" is specified, an order is established between X and Y, but not between them and the remaining values. However, participants that provided this kind of preference had in mind that they specified that X and Y are preferred to all the other values, and they could not see how this preference was taken into account in the presented choices.
- *Establishing a preference indifference without specifying preferred values.* Similarly to above, some participants stated that they are indifferent to X and Y, but did not specify that these are preferred values. Therefore, they also could not see how this preference was taken into account in the presented choices.
- *Providing a hard constraint that is not compatible with most of the models.* One participant provided a hard constraint that caused most of the models (which satisfy her other preferences) to be discarded, and received choices that did not satisfy her other preferences. As she did not pay attention to this constraint when analysing options, she did not agree with the choices made.

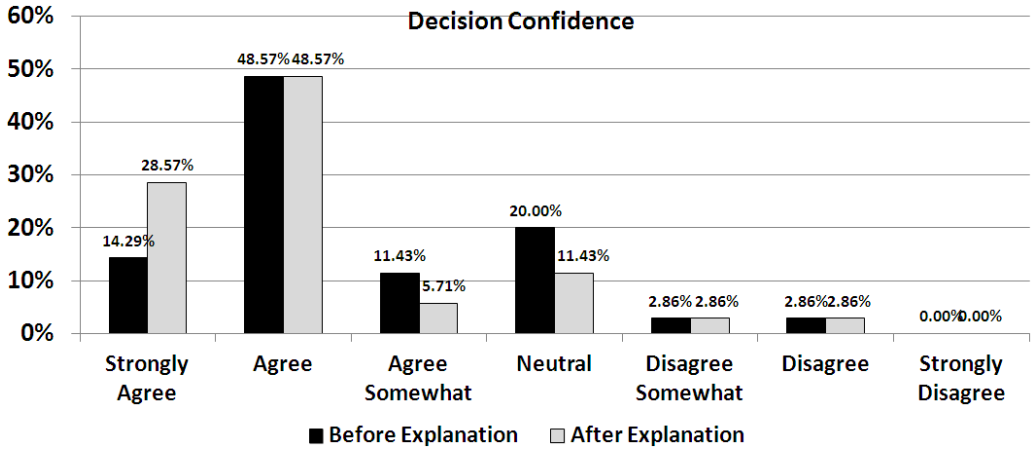
An interesting situation is that some participants did not (strongly) agree with the decision, but they said they understood it. These participants realised that they forgot to specify something in their preferences as the first choices had undesired characteristics, which they did not mention as a preference.



10.3(a): Choice Quality.



10.3(b): Trust in Choice.



10.3(c): Decision Confidence.

Figure 10.3: Choice Evaluation and Explanation Impact.

Measurement	Before Explanation		After Explanation		Wilcoxon Test	p-value
	M	SD	M	SD		
Choice Quality	5.97	1.12	6.09	1.15	$W(34) = 3.50$	0.102
Trust in Choice	5.86	1.09	6.03	1.01	$W(34) = 2.50$	0.084
Decision Confidence	5.43	1.22	5.80	1.23	$W(34) = 0.00$	0.006

Table 10.4: Choice evaluation and explanation impact measurements.

Even though our technique received high levels on the evaluation of choices made, only 26.47% of the participants would choose the option chosen by the technique, and 52.54% would choose one of the up to five selected options.

After receiving explanations generated with our approach, participants had to answer these same questions in order to evaluate the impact of the explanations. It can be seen some participants changed their evaluation from less positive to more positive rates, showing that the explanations tend to increase the choice quality perceived by users, and the trust and decision confidence on choices made. A Wilcoxon Signed Ranks test was conducted to compare these three measurements before and after explanations, and it showed that the explanations have a significant impact only on the decision confidence. Therefore, *we cannot reject hypotheses* $H1_0$ *and* $H2_0$*, but we reject hypothesis* $H3_0$. We summarise the results in Table 10.4.

Some participants provided only a few preferences over boolean attributes, which split the available mobiles phones into two groups: those that satisfy these preferences, and those that do not. In these cases, the explanations just expose this situation, stating that a random choice was made between options that satisfy all preferences, which explains the choice but it is not very helpful, thus most of the participants in this case did not change their evaluation rates. There was only one case in which a participant decreased her rates on these measurements, showing one shortcoming of our approach. This participant specified a preference for an attribute, and later she added a don't care preference related to this same attribute. Therefore, the decision support models (PSM and OAPM) took that preference into account, but the decision function ignored it because of the don't care preference. However, some of the explanations took into account the decision models and mentioned that attribute, and the participant complained that she stated she does not care about it.

Finally, participants had also to evaluate whether they could understand why the choice was made based on the provided explanations (*transparency*) and, as it can be seen in Figure 10.4, most of the participants agreed with that. We observed that participants expected that important attributes related to unsatisfied preferences should always be mentioned, which was not always the case. The average and standard deviation of the transparency measurement are $M = 6.20$ and $SD = 1.21$, respectively.

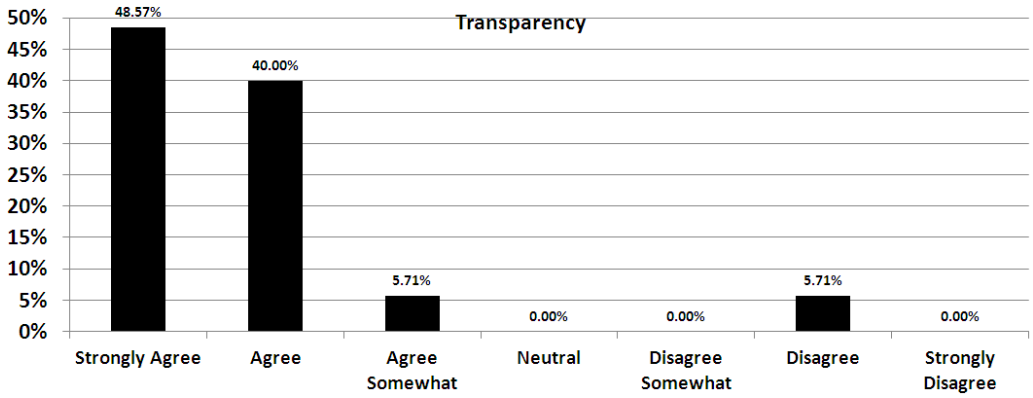


Figure 10.4: Explanation Impact — Transparency.

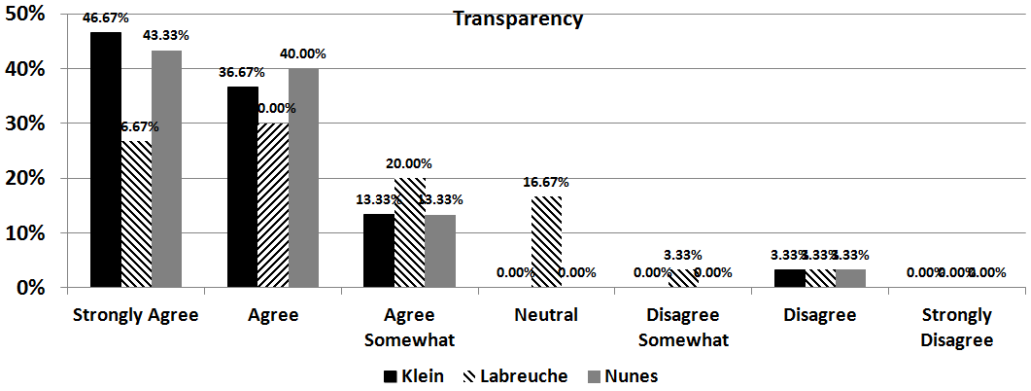
10.2.3 Explanation Comparison

After analysing the explanation impact, we compare explanations generated by different techniques. In order to do so, participants made a side-by-side comparison of three techniques, including ours, evaluating them with respect to the same criteria as above. Klein and Shortliffe's and Labreuche's approaches receive a utility function as input to generate an explanation that justifies why an option is better than another. As this comparison is made in a pairwise fashion, we can use the values of our cost function as input of these approaches.

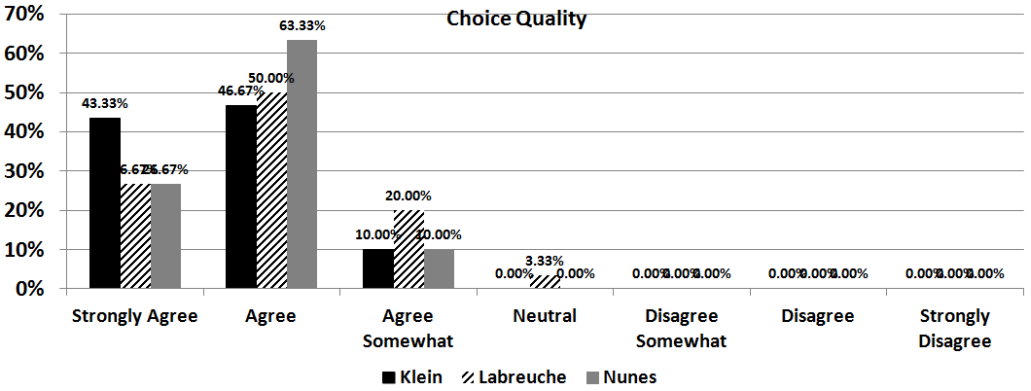
It is important to highlight that we established a timeout of $2min$ for each approach to generate explanations, as one of the possible explanations of Labreuche's approach is associated with a branch-and-bound algorithm, which can take a considerable amount of time to run (the algorithms of the other approaches run in polynomial time). This timeout was tested with a pilot study, and it showed that giving more time to the execution of the algorithm would make the participants lose engagement in the study. When this timeout is reached, all the explanations generated up to this moment are shown. While other approaches always executed in less than $2min$, there were five cases in which Labreuche's approach presented no results (which were discarded for comparing the different explanation techniques), and ten other cases that it presented six or less explanations. This already points out a limitation of Labreuche's approach.

We begin by presenting the collected data during this step, and we will later discuss relevant points we observed during the execution of this part of the study. The results of the comparison of transparency, choice quality, trust in choice and decision confidence among the three explanation approaches are shown in Figures 10.5 and 10.6.

First, we noticed that participants focused on the first selected options (this observation is also valid for the two steps above), and sometimes they did not even



10.5(a): Transparency.



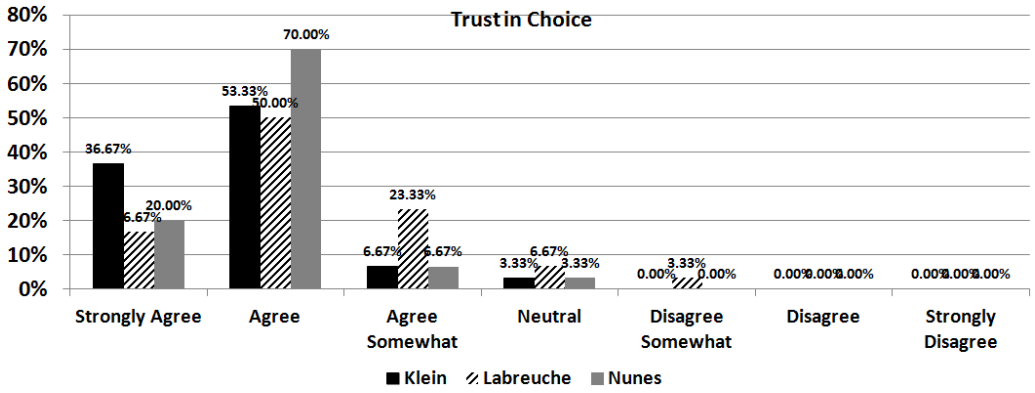
10.5(b): Choice Quality.

Figure 10.5: Explanation Comparison (I).

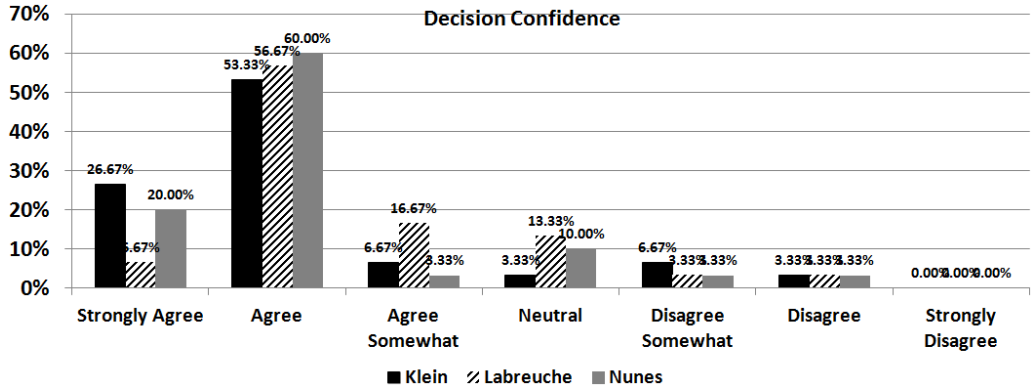
open the view in which all options with explanations are shown. In some cases, when participants had a specific preferred model, they searched for this model to verify why it was not chosen. Therefore, participants may have answered questions based on a subset of explanations.

The explanation approach that received the best scores on average was Klein and Shortliffe's. This approach provides only one type of explanation, which selects a subset of relevant attributes based on a threshold. As a consequence, their approach basically exposes the main positive (and possibly) negative aspects of options, helping the user to confirm that the choice made is right or wrong, but it does not actually explain the choice. In some cases, the explanation can be quite long, as the threshold may include many attributes.

The worst rates were given to Labreuche's approach. Based on our observations, this is mainly due to the complexity of their approach. In many cases the explanation given follows the *invert* pattern, which indicates the relationship of important pros and not important cons. Participants read these explanations more than once to understand its content, and there were reports that the explanation is complex and too long. Additionally, there were cases in which the domination



10.6(a): Trust in Choice.



10.6(b): Decision Confidence.

Figure 10.6: Explanation Comparison (II).

pattern was presented, and participants usually gave lower scores for explanations involving many situations of domination. What happened is that in these cases options were dominated because of the participants' preferences, and not actually because they were worse in all aspects, and this was not considered an appropriate explanation by the participants. For example, if a participant states that she wants a mobile phone with 32GB of internal memory, according to her preferences a mobile phone with 64GB is "worse" than one with 32GB, but in practice it is not.

This domination problem was also identified in our approach, which is one of the reasons why our approach received a few scores lower than Klein and Shortliffe's approach, and an example of this situation is shown in the *Domination* row of Table 10.5. Another problem we have identified is associated with the explanation for the chosen option, which highlights the key aspects that caused this option to be chosen. Many participants interpreted that the mentioned attributes in this explanation (mainly when there was only one attribute) were the only ones taken into account, such as the case shown in Table 10.5 (*Chosen Option Explanation* row). Because of these issues, which indicate possible improvements to be done, our approach was rated with lower scores in these two situations

Situation	Klein	Nunes
Domination	The attributes front-facing camera resolution (MP), and touch screen provide the most compelling reason to prefer mobile phone X over mobile phone Y.	There is no reason to choose option mobile phone Y, as option mobile phone X is better than it in all aspects.
Chosen Option Explanation		Option mobile phone X was chosen because of its brand.
More concise		Option mobile phone X was chosen because of its camera resolution (MP), and price.
	While height (cm) is a compelling reason to prefer mobile phone W over mobile phone X, price (US$) is a compelling reason for not choosing it.	Even though option mobile phone W provides better height (cm) than the chosen option, it has a worse price (US$).
	While height (cm) is a compelling reason to prefer mobile phone Y over mobile phone X, camera resolution (MP), touch screen, and price (US$) is a compelling reason for not choosing it.	Option mobile phone Y was rejected because of its touch screen and camera resolution (MP).
	While height (cm) is a compelling reason to prefer mobile phone Z over mobile phone X, downloadable games, camera resolution (MP), touch screen, and price (US$) is a compelling reason for not choosing it.	Option mobile phone Z was rejected because it does not satisfy constraints associated with downloadable games.
Different Text	The attribute QWERTY keyboard provides the most compelling reason to prefer mobile phone X over mobile phone Y.	Option mobile phone Y was rejected because of its QWERTY keyboard.

Table 10.5: Examples illustrating the main differences between Klein and Shortliffe's approach and ours.

(for 8 participants). However, *in general our approach received similar scores to their approach, and sometimes even higher* (for 8 participants as well), which are associated with situations in which Klein and Shortliffe's threshold is not adequate — as exemplified in the *More concise* of Table 10.5. We also noticed that the "*cost-benefit relationship*" causes a positive impact on the explanations, and when it was given to participants, they gave higher scores to our approach than to the others.

Finally, we make a comment on the text associated with the explanations — we implemented each approach following the text suggested by their respective authors. Explanation approaches focus on the algorithms that select parameters of templates to build explanation sentences. Nevertheless, in some cases Klein and Shortliffe's approach and ours selected the same attributes to present to participants, and their only difference was the text associated with it, as shown in row *Different Text* of Table 10.5, and the rates given to them were different. One participant also complained that Labreuche's approach mentions the mobile phone name too many times in the explanation. Issues related to the text of the sentences are easy to be

Measurement	Klein		Labreuche		Nunes		Friedman's Test	p-value
	M	SD	M	SD	M	SD		
Transparency	6.20	1.06	5.50	1.33	6.17	1.05	$\chi^2(2) = 2.7578$	0.0161
Choice Quality	6.33	0.66	6.00	0.79	6.17	0.59	$\chi^2(2) = 2.6958$	0.0193
Trust in Choice	6.23	0.73	5.70	0.95	6.07	0.64	$\chi^2(2) = 3.3428$	0.0024
Decision Confidence	5.80	1.27	5.40	1.13	5.73	1.20	$\chi^2(2) = 2.6817$	0.0201

Table 10.6: Explanation comparison measurements.

tackled, and must be taken into account as they are related to how users perceive the quality of the explanations.

In summary, Klein and Shortliffe's explanations are generally good, as showing option pros and cons is in general helpful, but there are cases in which a more specific explanation is better. This is what Labreuche's approach and ours aim to do, but they generate inadequate explanations in some cases. While Labreuche's approach turned explanations too complex for the participants, our approach managed to provide good explanations for them, with some exceptions, which were listed above.

In order to test whether the difference among the different explanation techniques is statistically significant, we used the Friedman's test. As it can be seen in Table 10.6, all the measurements differ significantly across the three explanation approaches. Therefore, *we can reject hypotheses* $H4_0$*,* $H5_0$*,* $H6_0$ *and* $H7_0$.

Given that we have a significant difference for the measurements, we further performed the post-hoc tests of Wilcoxon-Nemenyi-McDonald-Thompson, which shows us that the differences are due to the following.

- **Transparency**
  - Klein and Labreuche (p-value= 0.0161)
  - Nunes and Labreuche (p-value= 0.0255)
- **Choice Quality**
  - Klein and Labreuche (p-value= 0.0192)
- **Trust in Choice**
  - Klein and Labreuche (p-value= 0.0024)
- **Decision Confidence**
  - Klein and Labreuche (p-value= 0.0200)
  - Nunes and Labreuche (p-value= 0.0268)

Therefore, we conclude that (i) Klein and Shortliffe's approach provides a significant improvement for all measurements with respect to Labreuche's

approach; (ii) our approach provides a significant improvement for transparency and decision confidence with respect to Labreuche's approach; (iii) Klein and Shortliffe's approach and ours have no significant difference for all measurements.

10.2.4 Approach Evaluation

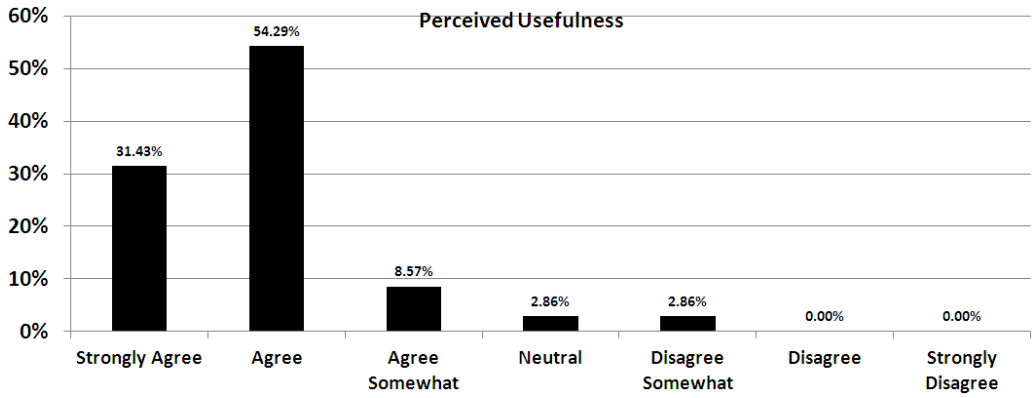
In the last step of the experiment, participants had to answer questions whose goal is to evaluate the whole approach: the experience while providing preferences, the choices made, and the explanations given. Figures 10.7 and 10.8 depict the results obtained, showing that a representative amount of participants answered (strongly or somewhat) agree for the *perceived usefulness* (94.29%), *intention to return* (97.14%), *enjoyment* (97.14%) and *satisfaction* (94.28%). There were participants that declared that they indeed needed a mobile phone and they were happy with the system and the recommended choices. Table 10.7 shows the average and standard deviation for all measurements.

Two of the measurements do not follow this case: *intention to purchase* and *intention to save effort*. Although many participants stated that the choice quality is good (first five options), they were not sure or disagreed that the chosen option is the best for them. Consequently, it may justify why 17.14% of the participants answered neutral or disagree with respect to the intention to purchase. In addition, this measurement depends greatly on purchase habits (e.g. impulsive vs. careful), being also a reason for the lower intention to purchase average. In fact, we included this measure in our study as it is part of Chen and Pu's framework (Chen and Pu 2010), but given that there are many variables that influence this measure (e.g. purchase habits), it may have a higher or lower score according to the personality of the participant and not her opinion with respect to the choice made.

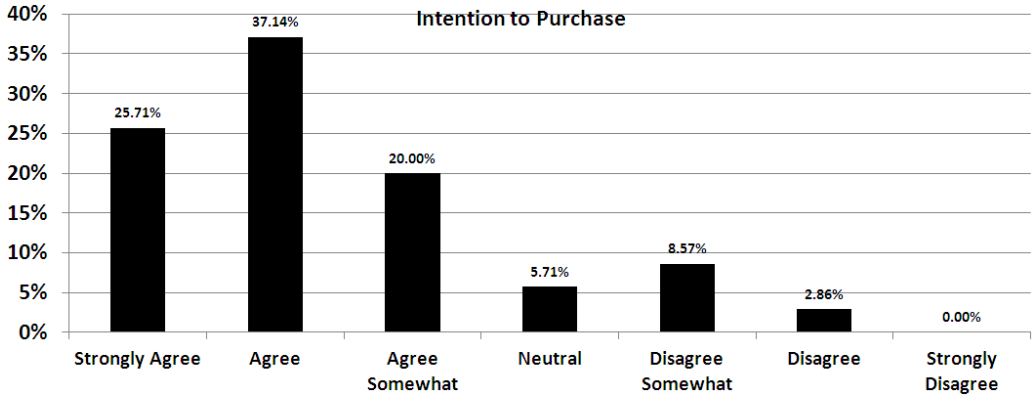
Regarding the intention to save effort, some participants claimed that even though using the system may help them with their choice, it would require more effort — but they are willing to use it anyway, because they believe it can help them to make a better choice than that they would do without support. One participant stated: "*if I wanted to buy an expensive, delicate, unique, etc. product, I would use the system because it is more sophisticated than I would choose myself.*"

Measurement	Average	Standard Deviation
Perceived usefulness	6.09	0.89
Intention to purchase	5.57	1.33
Intention to return	6.23	1.06
Intention to save effort	5.94	1.19
Enjoyment	5.83	0.92
Satisfaction	6.06	0.84

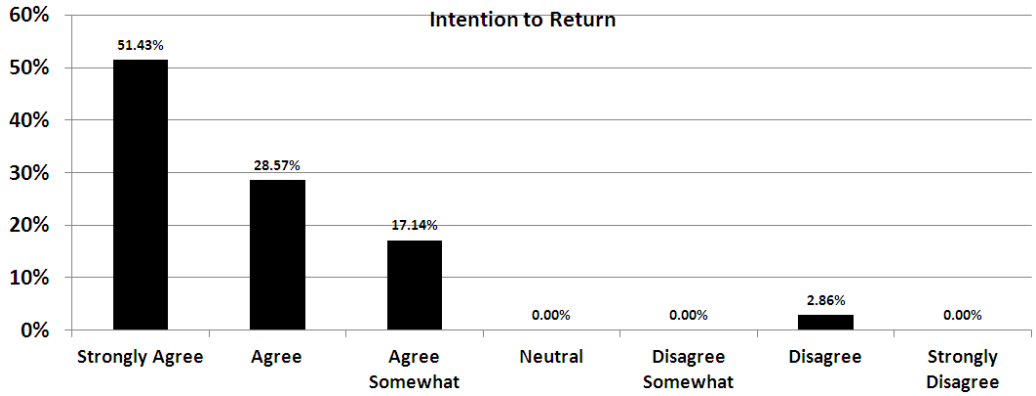
Table 10.7: Approach evaluation measurements.



10.7(a): Perceived usefulness.

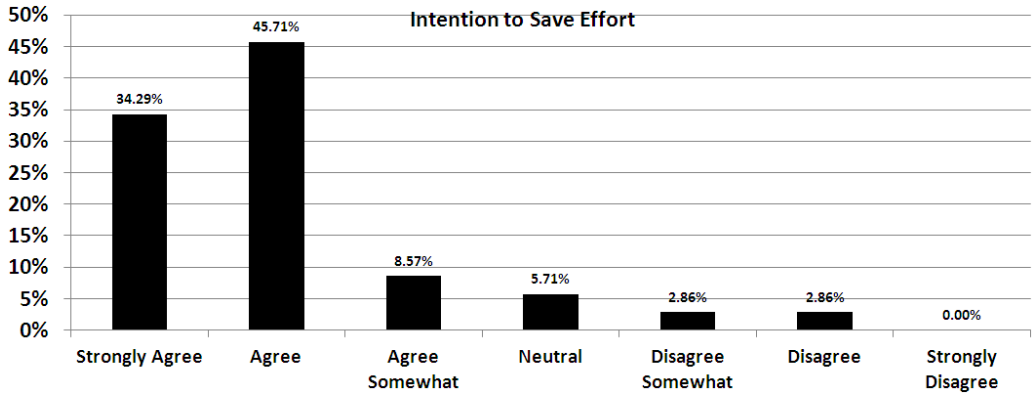


10.7(b): Intention to Purchase.

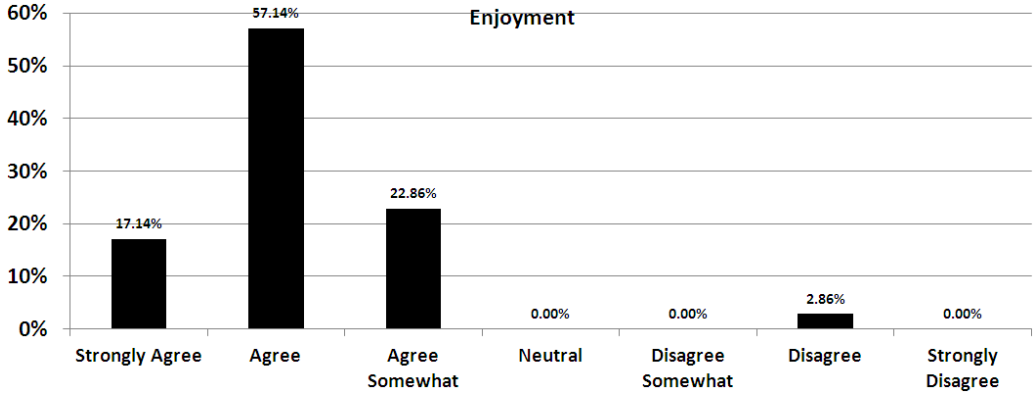


10.7(c): Intention to Return.

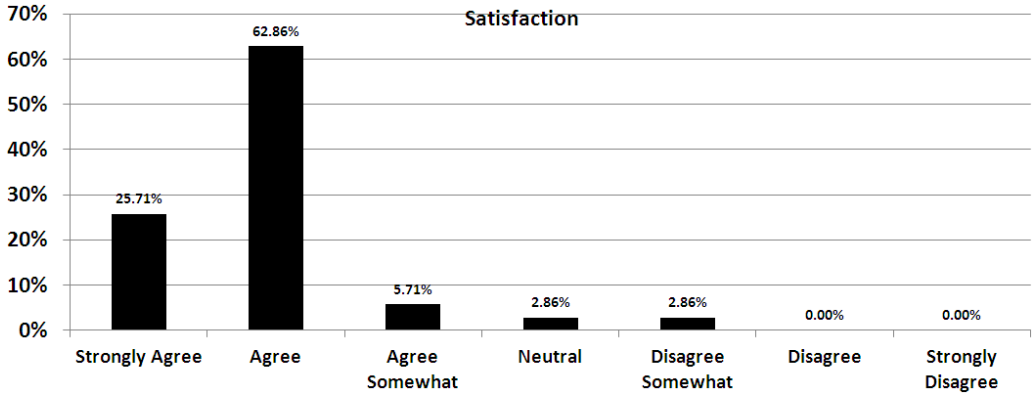
Figure 10.7: Approach Evaluation (I).



10.8(a): Intention to save effort in next visit.



10.8(b): Enjoyment.



10.8(c): Satisfaction.

Figure 10.8: Approach Evaluation (II).

10.3 Threats to Validity

In this section, we discuss threats to the validity of this study. We have identified three possible threats, which are listed below.

Construct Validity. In order to evaluate how different explanation approaches impact the trust in choice, participants had to answer a question related to this measure for each of the investigated approaches. However, we identified that participants may have answered the question related to trust from two perspectives: (i) trust in the decision maker (for choice evaluation and explanation impact); (ii) trust in the explanation given (for explanation comparison). Therefore, the trust in choice measure analysed for explanation comparison may be not related to the combination of the choice and explanation, as it was intended. Moreover, we observed that it was hard for the participants to isolate the three explanations to evaluate how confident they were in the decision when receiving each explanation. This problem can be tackled by performing another study using a between-subjects design, but it has the disadvantage that different participants may have different perceptions for the rates.

Internal Validity. Most of the participants had previous knowledge about the available mobile phones. As a consequence, when they answered questions to evaluate the choice quality, trust in choice and decision confidence, they may have used this knowledge. Therefore, the impact of the explanation can be lower than in the situation in which participants do not know options of the available set, such as when people search for hotels.

External Validity. The fact that our user study involved participants (i) of single geographic location (i.e. Brazil) and (ii) that are volunteers is a threat to the generalisation of the results of our study.

10.4 Final Remarks

In this chapter, we presented a user study, which was performed to evaluate different aspects of our approach: the preference language, the decision making technique and the explanation generation approach. The study consisted of allowing participants to interact with an application in which they had to (i) provide their preferences according to our language; (ii) analyse a choice made based on their preferences; (iii) analyse different explanations that justify this choice. Each of these steps also involves answering questions to evaluate our approach.

The results of our study show that our preference language is expressive enough for users to provide their preferences, but for some of them this process demands too much effort. Even though our approach does not include an interface for specifying preferences, participants considered the one used in the study easy to use. Our decision making technique was also positively evaluated, achieving high rates for choice quality, trust in choice and decision confidence. Moreover, explanations generated with our technique significantly increase the decision confidence. Our explanation technique was also compared with two main existing approaches, and results showed that our technique and also Klein and Shortliffe's approach are better than Labreuche's approach with respect to transparency and trust in choice. Although the difference between our explanation approach and that proposed by Klein and Shortliffe is not statistically significant — they are indeed considered equally good for many participants — the former outperforms the latter in some situations, at the cost of providing worse explanations in others.

Our study allowed us to identify shortcomings in the three aspects of our approach (language, decision making and explanation), which will help us to further improve our work. Finally, other studies can be conducted to refine the results of this study, tackling identified threats to its validity.