An Empirical Evaluation of Interactive Visualizations for Preferential Choice

Jeanette Bautista and Giuseppe Carenini
Department of Computer Science
University of British Columbia
201-2366 Main Mall
Vancouver BC V6T 1Z4
bautista, carenini@cs.ubc.ca

ABSTRACT

Many critical decisions for individuals and organizations are often framed as preferential choices: the process of selecting the best option out of a set of alternatives. This paper presents a task-based empirical evaluation of ValueCharts, a set of interactive visualization techniques to support preferential choice. The design of our study is grounded in a comprehensive task model and we measure both task performance and insights. In the experiment, we not only tested the overall usefulness and effectiveness of ValueCharts, but we also assessed the differences between two versions of ValueCharts, a horizontal and a vertical one. The outcome of our study is that ValueCharts seem very effective in supporting preferential choice and the vertical version appears to be more effective than the horizontal one.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Graphical user interfaces (GUI)*; I.3.6 [**Computer Graphics**]: Methodologies and Techniques—*Interaction Techniques*; H.4.8 [**Information Systems Applications**]: Types of Systems—*Decision Support*

General Terms

Design, Experimentation, Human Factors

Keywords

Visualization techniques, preferential choice, empirical evaluation, user studies

1. INTRODUCTION

Developing effective interactive visualization interfaces requires a possibly long process of iterative design, in which task analysis, analytical evaluation and user studies are successively applied. We have followed this methodology in the development of an interactive visualization framework to support preferential choice: the process of selecting the best option out of a set of alternatives. Preferential choice has been extensively studied in decision theory as many critical decisions for individuals and organizations are often framed as preferential choices. For instance, selecting a house to rent or buy, deciding who to hire, selecting the location of a new store or deciding where to spend your next vacation are all examples of preferential choices. When people are faced with such decisions, they look for an option that dominates all the others on all aspects they care about (objectives in decision theory). However, such an option often does not exist. For instance, when selecting a house within a specified price range, you may find one that is situated at the ideal location but does not have all the amenities you seek. In this case you will have to consider the tradeoffs. People are generally not very effective at considering tradeoffs among objectives, and require support to make this process easier [9].

According to prescriptive decision theory, effective preferential choice should include the following three distinct interwoven phases. First, in the model construction phase, the decision maker (DM) builds her decision model based on her objectives: what objectives are important to her, the degree of importance of each objective, and her preferences for each objective outcome. Secondly, in the inspection phase, the DM analyzes her preference model as applied to a set of alternatives. Finally, in sensitivity analysis, the DM has the ability to answer "what if" questions, such as "if we make a slight change in one or more aspects of the model, does it effect the optimal decision?" [9].

In the development of interactive tools for preferential choice, we argue that full support for - and fluid interaction between - all three phases are essential in making good decisions.

In [8] we presented ValueCharts (VC), a set of interactive visualization techniques to support preferential choice. VC in its original form was designed by mainly focusing on supporting the model inspection phase. Furthermore, the design of the interface relied on a rather simple task analysis exclusively based on decision theory.

In [5], we described the second major iteration in the development of VC. We presented the Preferential choice Visualization Integrated Task model (PVIT): a much more sophisticated compilation of domain-independent tasks that considers all aspects of preferential choice, some new ideas from decision theory, and more importantly, an integration

AVI '08, 28-30 May , 2008, Napoli, Italy.

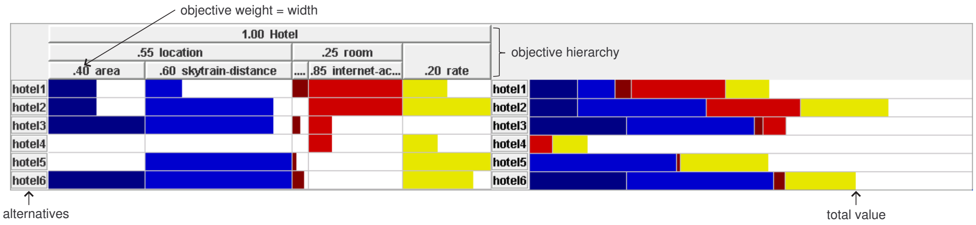

Figure 1: Horizontal ValueCharts - VC+H

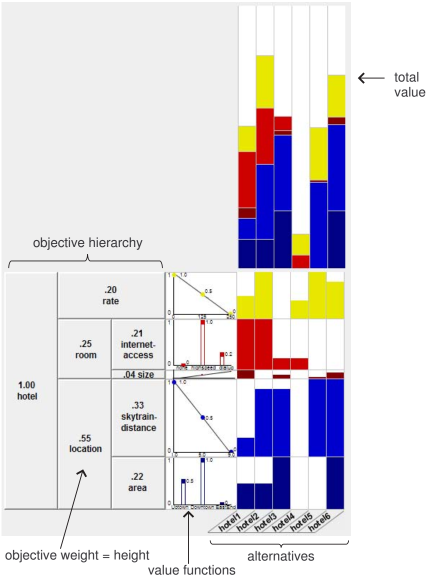

Figure 2: Vertical ValueCharts - VC+V

of task frameworks from the area of Information Visualization (InfoVis). For a detailed account of how tasks from literature in InfoVis and Decision Theory were integrated into our model, please see [5].

A new version of VC, called VC+, was developed to effectively support all the tasks included in PVIT (See [5] for the rationale behind the redesign). We also used the PVIT model to compare analytically VC+ with other existing tools for preferential choice [3, 2, 6]. VC+ appears to clearly dominate all its competitors (30% more effective). We developed two alternative versions of VC+. In VC+H the information is displayed horizontally, while in VC+V the information is displayed vertically. The PVIT model suggests that VC+V should be more effective than VC+H.

In this paper, we present our extensive empirical testing of VC+. In our evaluation approach we follow [16]. In particular, as advocated by Plaisant, we give subjects real problems, we ground the evaluation in a comprehensive task model and we measure both task performance and insights.

Since the analytical evaluation indicated that the other tools fare considerably worse than VC+ in supporting the tasks of the PVIT model, we gave low priority to a comparison study. Instead, we performed a comprehensive user study, grounded in the PVIT, to verify the usefulness and effectiveness of VC+, as well as focused on the assessment of the differences between the horizontal (VC+H) and the vertical (VC+V) versions.

The key contribution of this paper is that we applied several distinct approaches in order to achieve a more comprehensive assessment. First, we performed a quantitative, controlled usability study to see how effectively users performed the low-level tasks of the PVIT. Second, we qualitatively observed subjects using the tool in a real decision-making context of their choice (out of three possible ones). The subjects then answered a number of questions regarding their experience with VC+ in the decision-making process. In addition, we attempted to measure the users' insights in the decision problem. And finally, we used interaction logging throughout the experiment for further study. By triangulation of methods, we aimed to more fully understand the DM's experience in using VC+ (and VC+H versus VC+V) to perform preferential choice.

As a preview of the paper, we first briefly summarize VC+ and the PVIT model. We then describe our evaluation methodology and experimental design. Next, we present the controlled experiment on the PVIT low-level tasks. Finally, we describe the exploratory study to assess the effectiveness of VC+ in terms of user experience and insights.

2. VALUECHARTS+ AND THE PVIT MODEL

2.1 ValueCharts+ (VC+)

We developed two variations of VC+. In VC+H the information is displayed horizontally (Figure 1), while in VC+V the information is displayed vertically (Figure 2). We will describe the general features of VC+ by referring to the vertical version and then we will examine the differences between the two versions.

VC+ is a set of interactive visualization techniques for preferential choice. It supports the DM in the construction, inspection and sensitivity analysis of a DM's preference model as an Additive Multiattribute Value Function (AMVF) [1]. In an AMVF the DM's objectives are hierarchically organized. In VC+ this hierarchy is displayed as an exploded stacked-bar, see Figure 2 left-bottom quadrant. The vertical height of each row indicates the relative weight assigned to each objective (e.g., *size* is much less important than *internet-access*). Each column represents an alternative, thus each cell portrays an objective corresponding to an alternative (bottom-right quadrant). The amount of filled color relative to cell size depicts the alternative's preference

[1] For a detailed description of AMVF and VC+, see also [8, 5]

value of the particular objective (e.g., the *rate* for *hotel4* is bad, *hotel3* is worst). The values are then accumulated and presented in a separate display in the form of vertical stacked bars, displaying the resulting score of each alternative (top-right quadrant).

Several interactive techniques are available in VC+ to further enable the inspection and sensitivity analysis of the preference model. For instance, center-clicking on an alternative label displays the corresponding domain values. Double-clicking on the row heading ranks the alternatives according to how valuable they are with respect to the corresponding objective. As an example of sensitivity analysis, an objective weight can be changed by sliding the row headings to the desired weight.

The two versions of VC+ we have developed are informationally equivalent. They differ only in how the same information is displayed and in how it can be accessed. As shown in Figure 2, besides the different orientation the main difference between VC+V and VC+H is that VC+V allows for persistent display of the AMVF's component value functions (Figure 2 center bottom-half). In an AMVF, there is one component value function for each objective, and it specifies how valuable different levels of the corresponding objective are to the DM. For instance, the lower the *skytrain distance* the better. As discussed in [5] including these functions persistently in VC+H would be visually misleading and possibly confusing, so in VC+H component value functions are only accessible on demand.

At first glance, it appears that offering a persistent view on the component value function presents only advantages. The DM should be able to more effectively inspect tradeoffs among objectives as the range of their levels is readily visible and objective names are also more readable, because the label width is now mainly affected by the depth of the tree instead of by the number of objectives. Yet, since the functions do take up some screen real estate, they can become less readable and useful if the number of objectives increases, and more importantly making them permanently visible requires a vertical orientation that may negatively affect some PVIT tasks.

An important goal of the user study presented in this paper is to clarify whether the different orientation and persistent component value functions affect the DM performance in using VC+V versus VC+H.

2.2 The Preferential choice Visualization Integrated Task model (PVIT)

The Preferential choice Visualization Integrated Task model [5] is a framework for the design and evaluation of information visualizations for preferential choice (See Figure 3). The PVIT model starts top-down from the task that defines preferential choice: *to select the best alternative.* The first decomposition is into the three main phases of preferential choice: *construction*, *inspection*, and *sensitivity analysis* of the preference model applied to the set of alternatives.

The next level below incorporates two task taxonomies from the area of Information Visualization. The first is a classic: Ben Shneiderman's task by data type taxonomy (TTT) [18], which includes the tasks from his information-seeking mantra of "*overview first*, *zoom and filter*, then *details on demand*", as well as *relate*, *history*, and *extract.* In more recent literature [1], Amar and Stasko recognize the need for information visualizations to not only support representation of data, as the TTT does well, but also facilitate higher-level analytical tasks such as decision-making and learning. To bridge what they call "analytical gaps" (the gaps between representation and analysis), we incorporated their high-level *Knowledge Tasks* (e.g., *concretize relationship*) into our PVIT model by expanding the *relate* task from the TTT (see center of Figure 3).

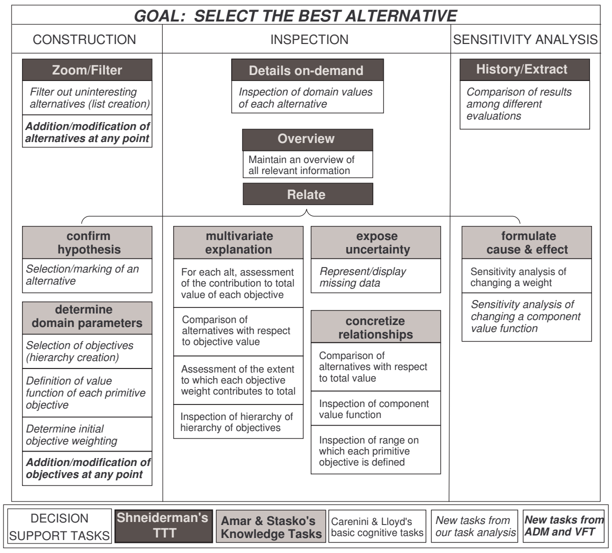
GOAL: SELECT THE BEST ALTERNATIVE
CONSTRUCTION
INSPECTION
SENSITIVITY ANALYSIS
Zoom/Filter
Filter out uninteresting alternatives (list creation)
Addition/modification of alternatives at any point
Details on-demand
Inspection of domain values of each alternative
Overview
Maintain an overview of all relevant information
Relate
History/Extract
Comparison of results among different evaluations
confirm hypothesis
Selection/marking of an alternative
determine domain parameters
Selection of objectives (hierarchy creation)
Definition of value function of each primitive objective
Determine initial objective weighting
Addition/modification of objectives at any point
multivariate explanation
For each alt, assessment of the contribution to total value of each objective
Comparison of alternatives with respect to objective value
Assessment of the extent to which each objective weight contributes to total
Inspection of hierarchy of hierarchy of objectives
expose uncertainty
Represent/display missing data
concretize relationships
Comparison of alternatives with respect to total value
Inspection of component value function
Inspection of range on which each primitive objective is defined
formulate cause & effect
Sensitivity analysis of changing a weight
Sensitivity analysis of changing a component value function
DECISION SUPPORT TASKS
Shneiderman's TTT
Amar & Stasko's Knowledge Tasks
Carenini & Lloyd's basic cognitive tasks
New tasks from our task analysis
New tasks from ADM and VFT

Figure 3: The PVIT model

In further refining PVIT into primitive tasks (leaf nodes), we go back to decision theory by first considering the original set of basic decision tasks for preferential choice proposed by Carenini and Lloyd [8]. Then, we augment this set so that all the generic knowledge tasks are instantiated and also by taking into account relatively recent ideas from decision theory (i.e., Value Focused Thinking [9]).

3. EVALUATION METHODOLOGY

Our evaluation methodology relies on [16], in which several guidelines for effective evaluation of interactive information visualizations are proposed. First, it is crucial that the empirical study matches tools with users, tasks, and real problems. In our study, we give subjects real preferential choices and we ensure the decision situation is of interest to the subject by letting her choose among three possible scenarios. Second, the evaluation should be grounded in a comprehensive task model. We follow this guideline by relying on the PVIT. Third, evaluation should not be limited to task performance but should also try to measure discovery and insights. Our evaluation comprises two parts. In Part A, we took a quantitative approach by performing a controlled usability study to see how users performed the primitive tasks of the PVIT. In Part B, we followed a more qualitative approach by observing subjects using the tool in a real decision-making context. In this second part of the study, we attempted to measure the users' insight in the decision problem.

Once the subjects had completed Part B, they filled out a questionnaire regarding their experience with VC+ in the decision-making process. Interaction logging were also collected throughout the experiment for further study.

Since we were comparing two versions of the same tool, we decided on a between-subjects experimental design to avoid the obvious learning effect that would come with a within-subjects design. Each subject was assigned to either VC+V or VC+H interface.

Subjects were recruited through the Reservax [2] online experiment reservation system. Prior to beginning the experiment, each subject read, signed and dated a consent form and filled out a pre-study questionnaire. 20 subjects, all students at UBC, of age ranging from late teens to 50+, agreed to spend 60 minutes with our experiment and receive $10 in compensation. Of the sample, 8 were male. All subjects were fairly computer proficient, ranging from 10 - 50+ hours per week. Each subject worked with only one VC+ interface: 10 subjects worked with VC+V and the other 10 worked with VC+H.

We found a good match in the grouping of the subjects in each treatment. Both groups had the same breakdown in sex and English proficiency, and the average computer use was very close. There was a slight difference in average age group: VC+V subjects were a younger group overall, half of them being less than 20 years old, and in the VC+H group, most subjects were in the 20-29 age group. All subjects had no previous exposure to formal decision analysis methods.

4. PART A: CONTROLLED STUDY

In this first part of the evaluation, we tested the hypothesis that the difference between the two orientations influences subject performance on a set of tasks form the PVIT model. In addition to the total-time-to-complete and correctness of tasks, we looked at each task individually.

4.1 Tasks

For this controlled study, we used data sets accompanied by scenarios: for training, we used the scenario of shopping for a used television set, and for testing we put the user in the situation of deciding on a hotel to stay at in Vancouver. It was assumed that the construction phase had already been completed (it is the same in both versions of VC+), and participants performed inspection tasks interspersed with sensitivity analysis tasks.

We considered the following four basic types of sensitivity analysis tasks (instances of the *formulate cause and effect* in the PVIT model): (i) What if [objx]'s weight is increased of k, and consequently [obj y]'s weight decreased of the same amount? (ii) What if [objx]'s weight is increased of k, and [all other n objectives] decreased of k/n? (iii) What if a component value function is changed in a numerical domain (e.g. money)? (vi) What if it is changed in categorical domain (e.g., neighborhood)?

We considered all nine primitive inspection tasks from the PVIT model. However, since four of the inspection tasks are implied by sensitivity analysis tasks (e.g. *Inspection of component value function* is implied when the user is asked to perform value function sensitivity analysis), we explicitly tested only the remaining five (see Table 1 for an example of these five tasks mapped to the house domain).

What are the top 3 alternatives according to total value?	List the 3 highest valued houses
For a specified alternative, which objective contributes to its total value the most?	For HouseX, which is its strongest factor according to your preferences?
What is the domain value of objective x for alternative y?	How many bathrooms are there in House1?
What is the best alternative when considering only objective x?	Which is the least expensive house?
What is the best outcome for a objective x?	What is the best bus-distance?

Table 1: Sample inspection tasks mapped to the House domain

4.2 Tutorial and Training

The PVIT construction tasks are assumed to help the user bridge the *Rationale* gap [1], i.e., they help the DM learn about the decision problem at hand, the model, and the decision analysis technique in general. So, although construction tasks were excluded from our evaluation, it was important to include the construction interface in the training.

The training session was performed on the TV domain. With the construction interface, the experimenter explained the objective hierarchy, the given alternative data, specifying value function, and objective weighting. After constructing the chart, the experimenter described the inspection interface in detail, covering all the types of tasks that the subject was to complete. The subject was then given a set of tasks to perform on the given model, which were task examples of the testing phase.

4.3 Procedure

After the training phase, each participant had an opportunity to ask questions for clarification before the testing phase began. Subjects were reminded that time and correctness were being measured, and that this time they did not have the opportunity to ask questions.

Once again, the experimenter walked the subject through the construction of the model (this time using the Hotel data set), but the testing did not start until after the VC+ view was in place. The subject was then given a set of tasks much like what they saw in the tutorial. The set was organized so that after each of the sensitivity analysis tasks, subjects completed a round of inspection tasks. In total, there were five rounds of inspection tasks (including one when the interface is first presented). Each subject performed each task, writing down the answer to applicable tasks that asked a question about the data.

4.4 Results

All subjects completed the procedure successfully. In terms of correctness of tasks performed, there was no significant difference between the two interfaces. In fact, there were very few mistakes made during testing and the average overall mean score for VC was 18.5, or 97.4% correct (participants scored 18.6 with VC+V and 18.4 with VC+H).

The high percentage of correctness does give us a good indication that subjects did well. However, we could not determine if the subjects performed well overall for time to complete tasks, since there is no benchmark to compare

[2] http://www.reservax.com/hciatubc/index.php, *HCI@UBC Subject Sign-up System*

against in this measure. Instead, we looked closely at these results to find an indication of whether one version of VC+ was better than the other for performing the tasks.

Our analysis follows a two-step approach that has been already successfully applied in Computational Linguistics [11], as well as in Human-Computer Interaction [14, 4], when two systems are compared on a relatively large number of tasks. First, you verify whether the performance of the two systems differ in a statistical significant way both across tasks and when performance for all tasks is aggregated (using the t-test). Then, you verify whether the two systems differ in a statistical significant way in the number of tasks in which one system is better than the other (using the Sign test [19]).

The mean time to complete all tasks was only slightly better for VC+V than VC+H for the training phase (19%). Although still non-significant, there was a more prominent difference seen in the testing phase (30%), in which subjects performed better with VC+V. When we broke down the evaluation by task, there were similarly no significant differences found.

Finally, in the second step of our analysis we determined, for each task, what interface the subjects performed better. VC+V performed better on all five inspection tasks and also performed better on three out of the four sensitivity analysis tasks. Then we applied a two-tailed Sign Test [19] to the obtained data (VC+V better 8 out of 9). This test measures the likelihood that the subjects performed better on one version over the other on m or more out of n independent measures under the null hypothesis that the two versions are equal. This test is insensitive to the magnitude of differences in each measure, noticing only which condition represents a better result. The outcome of this test is that overall subjects performed significantly better with VC+V in the testing phase ($p = 0.039$).

According to our analysis we can conclude that even though there are no significant differences in training time between the two interfaces, subjects work more efficiently with the vertical interface after the initial training.

In addition to these overall results, we looked closely at individual task results and interaction logs and found some interesting observations. For example, we were able to understand why there were no significant differences in time to complete value function sensitivity analysis tasks. In the VC+H subjects took extra time because they had to recall what the value function was and how to access the hidden display, whereas subjects did not experience this problem in VC+V. They did, however, take longer to interact with the smaller display, and some subjects ended up opening the on-demand view. Although the persistent display did not affect time to perform the task, we will see that it played a bigger part in overall decision-making (See Part B below). We also found problems in our design regardless of orientation. For example, we found that subjects had trouble with the pump function for sensitivity analysis of weights (see [5]) in both VC+H and VC+V. These and several observations by task will be taken into consideration with future design iterations of ValueCharts.

5. PART B: EXPLORATORY STUDY

In Part A of our evaluation, we looked closely at how subjects performed tasks that are important for effective analysis in decision-making. Because these tasks still need to be appropriately combined to lead to effective preferential choice, in Part B we attempt to more fully understand the DM's experience in using VC+ to perform preferential choice. To achieve this goal we observed subjects using the tool in a real decision-making context of their choice (out of three possible ones). After interacting with the system, subjects filled out a questionnaire regarding their experience with VC+ in the decision-making process.

5.1 Insight Characteristics

A primary purpose of visualization is to generate insight [7]. It has been argued that the generation of insights leads to a better understanding of the domain and problem situation, thus favoring better decisions. An effective visualization will aid the DM to see things that would otherwise go unnoticed, as well as enable her to view information about her preferences in a new light.

In our exploratory study we measure the amount of insight each subject gains from using VC+ for a particular decision-making scenario. We use the definition of insight provided in [17]:*an individual observation about the data by the participant, a unit of discovery.* In terms of our model, we consider the DM's preferences and weighting as part of the data observed.

Saraiya and North in [17] propose an evaluation protocol for insights based on a set of "characteristics of an insight". Although this set is assumed to work for other domains, it is accepted that it may require some adaptation.

Notice that Saraiya and North's study for evaluating insight is in a very specific and technical domain (i.e., microbiological and microarray data). And the subjects had extensive domain knowledge. In contrast, our study is less specific (subjects worked in different domains they could choose from), much less technical (e.g., house rental) and the subjects were not experts.

Based on these observations, in our study we applied some slight modifications and generalizations to Saraiya and North's original set "characteristics of an insight".

The following is our characterization of insight as applied to preferential choice:

- **Fact:** The actual finding about the data (e.g. "Samsung [cell phones] are the smallest")
- **Value:** How to measure each insight? We determined and coded the value of each insight from 1 - 3, whereas simple observations of domain value and top ranking (e.g. "cheapest place is in East Van") are fairly trivial, and more global observations regarding relationships and comparison (e.g. "more expensive phones have all the features") are more valuable.
- **Category:** Insights were grouped into several categories:
  - Simple fact: an alternative rank or identification of domain value e.g. "This phone is fairly light", "This phone is only [ranked] fourth for battery"
  - Sensitivity: how a change affects the results e.g. "This house again!", "Now this phone is third"
  - Realization of personal preferences: users often stated that they made a realization about their preferences e.g. "it makes sense, because I really like hiking and nature", "brand should be more important [to me]"

These categories were defined after the experiment, and the grouping closely lends to the value coding.

5.2 Domain Data Sets

In order to ensure that the users had the capability to determine insightful facts about the information presented to them, it was important that they had a genuine interest in the domain that was studied. The subjects were asked to choose one out of three different decision problems. Each of the decisions included a scenario in the following domains:

House Rental: data was taken loosely from current postings on AMS Rentsline, where any missing information was fabricated. General information, such as rent, location, type, etc, were consistently available, but other more detailed information (bedroom size by sq-ft) was often fabricated. The scenario is that the DM goes to school at UBC and would like to move off campus. It is assumed that the DM is only considering Point Grey, Kitsilano, Downtown, and EastEnd. The House Rental decision problem contained 13 objectives and 10 alternatives.

Cell Phone: data was taken from Rogers Video website, and there were only a few cases of missing information. The information was narrowed down to 17 primitive objectives, and anything the participant was looking for (i.e. text-messaging) was assumed to be a feature included in all phones. The scenario is that the DM is looking for a phone from Rogers Wireless based on a 3-year plan (as prices were quoted). The Cell Phone decision problem contained 17 objectives and 12 alternatives.

Tourism: in this situation, the data was taken from Tourism Vancouver Official Visitor's Guide. The alternatives were narrowed down to those listed as being Downtown, East End, West End, and North Van. The scenario is that the DM is looking to take a visiting friend to a local tourist attraction. Alternatives were further categorized as type (scenic, historic, etc.), and indoor/outdoor. Cost was assumed as average/adult. The Tourism decision problem contained 17 objectives and 12 alternatives.

We realize that one possible problem in this methodology is that the number of alternatives and objectives differ (which may affect the number of insights). We do, however, believe that the advantages dominate this possible disadvantage.

Each scenario was explained to the participants, and they were asked to choose which one they would like to work with.

5.3 Procedure

At the onset of Part B of the study, subjects have already undergone considerable training and practice from Part A. However, since the construction interface was not tested in the controlled study, experimenters worked with the subjects to build the initial decision model. Objectives were presented to them in a pre-existing hierarchy with all available factors, and were told to remove and rearrange as they pleased (additions were not allowed since data set was fixed and could not be extended).

To set their initial preference model, they were instructed to go through the list of objectives and set the value function of each one to reflect their true preferences. Default functions were provided, where typically linear continuous functions were given (i.e. positive for battery talk time, negative for price), and each discrete objective was set with a best, worst, and 0.5 for others. Finally, the subjects ranked the objectives with the SMARTER weighting technique [12]. Their resulting decision model was then presented with VC+. The subject was asked to use the interface to analyze the decision model, perform any sensitivity analysis changes as they see fit, and view any information that they required. They were instructed to work with the interface to make a decision about the data, where the decision could be to select one or more preferred alternatives.

Subjects were asked to "think aloud" as they analyzed the preference model, being sure to let the experimenter know anything interesting that they saw. Notes were taken by the experimenter, and interaction logging was turned on once VC+ was created.

The subject was asked to take as little or as much time as she needed in order to reach his decision. If she was finished quickly, the experimenter would probe, but end the session if she was satisfied with the decision. The time for the experiment (total of both Part A and B) was 60 minutes, and if subjects were approaching the 60 minute mark, they were warned by the experimenter but welcomed to stay until as long as the 75 minute mark.

At the end of the exercise the subject was asked what their decision was, and to keep that in mind when answering the post-experiment questionnaire.

5.4 Results

It appeared that every subject had a genuine interest in the domain that they chose (10 cell phone, 6 tourism, and 4 house). Overall, subjects were able to use the tool and conclude on a best decision.

Subjects went through the construction phase carefully. The time spent inspecting the interface (minus construction) ranged from 3-16 minutes. The number of insights ranged from 0 to 10.

5.4.1 Comparison between the two interfaces in terms of insights

Table 2 summarizes two measures of insight gained and usage time, illustrating the two different interfaces. It shows a) mean number of insights acquired, b) the mean sum of value for all insight occurrences, and c) the average total time each subject spent using the tool until they felt that they reached a decision.

VC+V	mean	sd	min	max
Count of insights	4.7	5.0	0	10
Total insight value	8.8	2.9	0	15
Total time	9.93	2.8	3.14	16.45

VC+H	mean	sd	min	max
Count of insights	3.3	6.2	0	10
Total insight value	5.9	3.2	0	15
Total time	8.69	5.3	3.03	12.65

Table 2: Insight Results

Statistical analysis indicates that there are no significant differences, despite the fact that there appeared to be a great difference in the mean insights and value (49% and 34% more, respectively). Because of these noteworthy differences we also measured effect sizes (the magnitude of the differences) to determine the practical significance of the differences. Cohen's d [10] provides a standardized measure of the mean difference between two treatments. In this measure $d > 0.8$ is considered to be a large effect, $0.8 > d > 0.5$ to be a medium effect, and $d < 0.5$ to be a small effect. We found the effect sizes of the insight count and insight value to be 0.40 and 0.51 respectively. So, although our results are not statistically significant, according to Cohen's criteria, using VC+V has a medium effect on the value of insights reported by our participants.

Since the evaluation method is more qualitative and subjective than quantitative, general comparison of the tendencies in the results is also appropriate. There were more insights counted for the vertical interface, which also fared better when value factor was considered. Looking more closely at the interaction logs reveal that subjects tended to perform more sensitivity analysis on VC+V, which in turn led to more insights on sensitivity. There were 89% more sensitivity analysis of value function performed on the vertical interface than the horizontal. We conclude that the reason for this is that the persistent view a) acts as a reminder of what the value function is and that it can be changed and b) is more inviting for users to directly manipulate value function. We hypothesize that there is a benefit from the persistent view of the component value functions, but may revisit the persistent sensitivity analysis technique in future iterations.

More time was spent on the vertical interface. In contrast to time measurement in Part A that we used to gauge performance of lower-level tasks, more time spent performing the overall task of making a decision can not be viewed as negative. In fact, the general trend was that the more time spent by the subject on the decision problem, more insights were reported.

It should be noted that, regardless of the interface, the results were very mixed. Some subjects did not have any insights, and some had many. The standard deviation was high overall (see Table 2). Individual differences were more apparent in this part (versus Part A) because subjects' personalities could affect the amount of insights reported (a challenge of the think-aloud technique [13, 15]). In addition, the possible varying level of interest in each subject's selected domains may contribute to this variance. Nonetheless, we believe that providing the subject with a selection of domains helped with degree of interest. A more extensive study might specify a single domain and recruit participants with a specific requirement (e.g. recruit participants who are in the market for a new cell phone, and plan to purchase or upgrade in the next month).

5.5 Post-study questionnaire

Following the exploratory study, we completed the session by asking the subject a number of open questions and having them fill out a post-experiment questionnaire. They were asked to answer each question by selecting the degree of agreement of the statement from 1 to 5 where 1 is strongly disagree and 5 is strongly agree. Some questions were specific to the exercise they performed in Part B, while others were about the overall experience of using VC+. This questionnaire provides information not only on differences between VC+V and VC+H, but also on the subjects' experience and satisfaction with VC+ in general.

All subjects were generally satisfied with the decision that they made ($\mu = 4.25, \sigma = 0.55$), although their level of confidence was slightly lower overall ($\mu = 3.95, \sigma = 0.76$). A closer look shows that 4 of 6 subjects who gave this a 3 or "neutral" rating had 3 or less insights. Subjects felt that VC+ was a good tool for learning about their preferences in the selected domain ($\mu = 3.95, \sigma = 0.69$). This was tied closely to insights as well, as we found a significant positive correlation between the rating of this question and insight. This analysis further supports the assumptions made in Part B that more (insights, time, interaction) is better.

Our subjects, who did not have any previous exposure to decision analysis methods, felt that they learned much about how to analyze their decision model ($\mu = 4.20, \sigma = 0.62$). We attribute this much to the construction interface that they were exposed to in training for Part A and working with building their decision model in Part B, since it represents tasks that support the higher-level analytical task of learning.

Overall VC+ was very well-received. All subjects thought that VC+ is useful, intuitive, easy to use and quick to learn. In particular, subjects rated the usefulness very high ($\mu = 4.40, \sigma = 0.50$), and strongly agreed that visualizing their preferences helps in their understanding of the decision ($\mu = 4.45, \sigma = 0.51$).

Details on the answers to the most informative questions in the post-study questionnaire are shown in Table 3.

	strongly disagree (1)	disagree (2)	neutral (3)	agree (4)	strongly agree (5)
I am satisfied with the decision I made					
VC+V	0	0	1	5	4
VC+H	0	0	0	8	2
I am confident about the decision I made					
VC+V	0	0	4	2	4
VC+H	0	0	2	7	1
I learned a great deal about my preferences in [selected domain]					
VC+V	0	1	1	6	2
VC+H	0	0	1	8	1
This is a useful tool for making decisions					
VC+V	0	0	0	5	5
VC+H	0	0	0	7	3
Visualizing my decision model helps me understand it more clearly					
VC+V	0	0	0	7	3
VC+H	0	0	0	4	6
I learned a great deal about how to analyze my decision model					
VC+V	0	0	0	6	4
VC+H	0	0	2	6	2

Table 3: Results of post-study questionnaire

6. CONCLUSIONS AND FUTURE WORK

We addressed some challenges of information visualization evaluation [16] in several manners. First and foremost, we developed and applied a taxonomy of tasks that represents a benchmark framework for design and evaluation of visualization techniques for preferential choice. In addition to a controlled experiment, we used a triangulation of methods that includes an exploratory study in which we matched users with real data in realistic scenarios and included a measure of insight.

We looked at ValueCharts in several angles with this evaluation focusing on comparing two versions with different orientations. First, we assessed how the subjects performed on the low-level tasks. On average the subjects performed well in correctness, varying to some degree in length of time spent to complete the tasks. In turn, when asked to perform the high-level task of making a decision with our tool, the subjects reported that they were quite satisfied with their decision. These results corroborate our claim that if an interface supports the lower level tasks of PVIT well, then the interface also will enable the higher level tasks of the model. We pruned our task model to focus more on the basic tasks of inspection and sensitivity analysis, which more directly support the higher level task of decision-making. Since subjects were generally satisfied with the construction phase as well, it added to the success of ValueCharts as tools for preferential choice.

Some of the evidence that we have collected suggest that the vertical and horizontal ValueCharts designs are not equivalent interfaces since a) the Sign test indicates that subjects perform better on the VC+V than VC+H on low level tasks, and b) VC+V has a medium effect on insight value as we explored how subjects performed the higher level task of decision making. However, the lack of statistical significance for the difference in insights (count and value) indicates the need for a larger experiment.

Nonetheless, our overall evaluation of ValueCharts Plus is very promising. Subjects rated our tool very high in usefulness, learning, and understanding.

In future iterations of the ValueCharts design, we would like to address some of the issues and observations that we discovered in these studies. We also plan to conduct a more extensive experiment using a larger pool of subjects and focusing on a single domain with participants screened for specific requirements (i.e. who are in the market of that particular domain). We will also consider some changes in our experimental procedure such as using other HCI experts to conduct the analytical evaluation. Additionally, we intend to conduct further studies of the construction interface.

7. REFERENCES

[1] R. Amar and J. Stasko. A Knowledge Task-based Framework for Design and Evaluation of Information Visualizations. In *Proceedings of InfoVis '04*, pages 143–150, Austin, TX, USA, 2004. IEEE Computer Society. Best Paper.

[2] N. V. Andrienko and G. L. Andrienko. Informed Spatial Decisions through Coordinated Views. *Information Visualization*, 2:270–285, 2003.

[3] T. Asahi, D. Turo, and B. Shneiderman. Visual Decision-Making: Using Treemaps for the Analytic Hierarchy Process. In *Proceedings of CHI '95*, pages 405–406, New York, NY, USA, 1995. ACM Press.

[4] R. Bade, F. Ritter, and B. Preim. Usability comparison of mouse-based interaction techniques for predictable 3d rotation. In *Smart Graphics*, pages 138–150, 2005.

[5] J. Bautista and G. Carenini. An integrated task-based framework for the design and evaluation of visualizations to support preferential choice. In *AVI '06: Proceedings of the working conference on Advanced visual interfaces*, pages 217–224, New York, NY, USA, 2006. ACM.

[6] V. Belton. VISA: Visual Interactive Sensitivity Analysis. SIMUL8 Corporation, Boston, MA, 2008.

[7] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc., 1999.

[8] G. Carenini and J. Lloyd. ValueCharts: Analyzing Linear Models Expressing Preferences and Evaluations. In *Proceedings of AVI '04*, pages 150–157, Gallipoli, Italy, 2004. ACM Press.

[9] R. T. Clemen. *Making Hard Decisions*. Duxbury Press, Belmont, CA, USA, 2nd edition, 1996.

[10] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences (2 ed.)*. Lawrence Earlbaum associates.

[11] B. DiEugenio, M. Glass, and M. J. Trolio. The DIAG experiments: Natural language generation for intelligent tutoring systems. In *The Second International Natural Language Generation Conference*.

[12] W. Edwards and F. H. Barron. SMARTS and SMARTER: Improved Simple Methods for Multiattribute Utility Measurement. *Organizational Behavior and Human Decision Processes*, 60(4):306–325, 1996.

[13] K. A. Ericsson and H. A. Simon. *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA, 1984.

[14] K. Hinckley, R. Pausch, D. Proffitt, J. Patten, and N. Kassell. Cooperative bimanual action. In *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 27–34, New York, NY, USA, 1997. ACM Press.

[15] J. Nielsen, T. Clemmensen, and C. Yssing. Getting access to what goes on in people's heads?: Reflections on the think-aloud technique. In *NordiCHI '02: Proceedings of the second Nordic conference on Human-computer interaction*, pages 101–110, New York, NY, USA, 2002. ACM Press.

[16] C. Plaisant. The challenge of information visualization evaluation. In *AVI '04: Proceedings of the working conference on Advanced visual interfaces*, pages 109–116, New York, NY,, 2004. ACM Press.

[17] P. Saraiya, C. North, and K. Duca. An evaluation of microarray visualization tools for biological insight. In *INFOVIS '04: Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'04)*, pages 1–8, Washington, DC, USA, 2004. IEEE Computer Society.

[18] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of VL '96*, page 336, Washington, DC, USA, 1996. IEEE Computer Society.

[19] S. Siegel and N. J. J. Castellan. *Nonparametric statistics for the behavioral sciences*. McGraw Hill, 1988.