DISCOUNT USABILITY EVALUATION OF AN ANALYTIC HIERARCHY PROCESS BASED CLINICAL COMPUTER APPLICATION

By

John Cooper

A CAPSTONE

Presented to the Oregon Health & Science University's

Department of Medical Informatics & Clinical Epidemiology

in partial fulfillment of

the requirements for the degree of

Masters of Biomedical Informatics

May 2005

Capstone Advisor

Karen Eden, Ph.D.

School of Medicine

Oregon Health & Science University

Certificate of Approval

This is to certify that the Capstone Project of

John H. Cooper

"Discount Usability Evaluation on an Analytic Hierarchy Process Based Clinical Computer Application"

Has been approved

[redacted]

Capstone Advisor

ACKNOWLEDGMENTS

I would like to thank a number of persons who were invaluable in their contributions to this research. First of all I want to recognize my advisor Karen Eden, without whom this entire project would have never been undertaken. From Karen's suggestions that helped me find a topic, to her coaching through the research design, to the lending of her research assistants, she has been an indispensable mentor. Along with Karen I also owe a great deal of thanks to the aforementioned research assistants, Deborah Rosenberg, Megan McClellan, and Jennifer "Dee Dee" Jimison. Not only did they assist by observing and interviewing subjects, they each contributed valuable feedback on the tools and methods I employed. Finally, I want to acknowledge the dedication of Nick Anderson of the University of Washington; Nick spent many hours he did not really have honing the decision support application for use in this project.

Outside the bounds of the academic world I wish to thank my wife, Zaida. Even as she brought home the bacon, she indulged me the time away from home needed to complete this project, never swaying in her support of this or any of my academic endeavors. I would also like to thank my daughter, Sonrisa, for the inspiration to be a student of both medical informatics and of life, and my parents, sister, brothers, and countless friends for helping me hang in there when times got tough.

TABLE OF CONTENTS

ABSTRACT

BACKGROUND: In this first-of-its-kind study I employed Discount Usability Engineering to evaluate and compare the usability of two versions of a clinical decision support application designed to help women make decisions concerning future births after having a cesarean.

METHODS: Eight women who had previously given birth were observed while engaging in the operation of limited prototypes of both application versions. They were encouraged to think aloud and were questioned to ascertain their understanding and opinions of each version.

RESULTS: Several important usability improvements were identified with low investment of time, resources, and funds. In particular this study demonstrated that the application appeared to be measuring women's birthing preferences more precisely than the women wanted.

CONCLUSION: Discount Usability Engineering was shown to be an effective method for discovery of meaningful usability improvements at minimal cost and effort.

Introduction

Overview

Usability is the idea that products need to be designed to cater to the average understanding and abilities of their user communities, as well as to a consensus of the users' preferred tastes. It is officially defined by the International Standards Organization in their publication 9241-11 (1998) as, "the extent to which the product can be used by specific users to achieve specific goals with effectiveness, efficiency, and satisfaction in a specific context of use."[1]

The usability study I describe in this paper was an adjunct project to ongoing research by Eden, Guise, Perrin, Dolan, Seshadri, Anderson, Rosenberg, McClelland, and Jimison (Eden *et al*), in the area of childbirth decisions after a cesarean[2]. The research of Eden *et al* is part of a broad-range effort to bring women who have borne a child via cesarean section directly into the decision making process for the type of delivery of their next pregnancy[3]. This research builds upon evidence that such women's preferences are not fully recognized either by themselves or by obstetric practitioners[4]. It also helps implement goals of a number of national and private health care organizations to enhance medical decision effectiveness[3].

The subject of the current research of Eden *et al* has been to assess the efficacy of a clinical computer application named the Preferences-Assessment Computer Module (PACM) to help women make birthing decisions[2]. In an earlier project, Eden, Anderson, Bhupatiraju, and Guise created PACM to enhance women's awareness of their true preferred choice between a repeat cesarean and a vaginal birth after cesarean (VBAC) should they undergo a successive pregnancy. They designed it for direct personal use by subject women. To accomplish this, it was developed using a decision support methodology known as the Analytic Hierarchy Process, specifically using a variant called Pairwise Comparison Method[5].

In an early phase of their current research Eden *et al* conducted a pilot effort to design the PACM application for optimum preference measurement. This effort resulted in two variations of the actual measuring mechanism within the application, but was inconclusive as to which variation was most effective. One, called the checkoff version, required users to answer questions by choosing from among a group of mutually exclusive user input fields using radio buttons. The other was fashioned as a sliding bar, consisting of a box that users could move horizontally along a wide line to a position of their choice. Both versions of the application displayed the exact same questions and employed the exact same algorithms for assessing user preferences[2].

The pilot study produced evidence that ability to measure patient preferences was adequate in both flavors of the application, provided that users understood and completed the measurements and accepted their output[2]. In other words, if users' answers accurately reflected their actual feelings, either version would provide representative calculation of their overall scores. Therefore any difference in effectiveness between the two should reflect differences in user responses and behavior toward the design, or in other words, differences in their respective usability. Presumably a more accepted and understood application would promote the validity of the research of Eden *et al.* As such, the purpose of my study was to aid Eden *et al* to determine which of the two versions was most usable -- most likely to be understood, accepted, and completed by the users. In addition the outcome of a usability study was expected to benefit the experience of research subjects.

Yet at the same time it was important to avoid substantially increasing either the cost or the time of the original research. In order to accomplish these goals I chose to conduct a usability evaluation between the two variations using the methodology known as Discount Usability Engineering.

Both Eden *et al* and I got more than we bargained for, but in a positive way. While my study showed that users might prefer the checkoff version slightly to the sliding bar, several issues were uncovered in both versions. Most importantly the study brought visibility to the varying ways the patients understood the preference measurements. The degree to which some women preferred repeat cesarean over VBAC births or vice versa may have been skewed by their avoidance or misunderstanding of some input entry values. Yet the discovery of this and other concerns was invaluable to Eden *et al* in improving the application itself and their research as a whole.

What is more, I discovered that clinical decision support applications had rarely ever undergone Discount Usability Engineering, while the Analytic Hierarchy Process had almost no record of usability evaluation of any kind. As a result, I hope to have pierced a frontier area of informatics research.

Background

Discount Usability Engineering

Discount Usability Engineering is a method of evaluating product usability. Nielsen, the father of Discount Usability Engineering, identifies five basic qualities of good usability:

- "Learnability" – the user finds the product simple to figure out and remember.
- "Efficiency" – the product easily helps the user be productive.
- "Memorability" – the occasional user can retain his/her experience with the product and reconstruct it easily.
- "Errors" [sic] – the product rarely provokes user mistakes, and the user can correct them without complication.
- "Satisfaction" – users enjoy or approve of how the product works[6].

According to Scholtz and the National Institute of Standards and Testing (NIST), usability enters a product through *usability engineering* (also known as human factors engineering (HFE)), which is a process of formatted development focused on user interfaces. Usability is ensured through *usability evaluation*, which is a component of HFE. Scholtz states that usability evaluation comes in three flavors:

- "User-centered" – non-expert product users engage typical features in expected manners while their experiences are observed and recorded.

- "Expert-based" – domain knowledgeable persons perform subject matter reviews, walkthroughs, and heuristic inspections.
- "Model-based" – usability is evaluated using models of human mental or physical behavior and perceptions[7].

All the above have their pros and cons. A major plus for user-centered evaluations is the involvement of the end user, but Scholtz cautions that their involvement can be costly, long-lasting, and may require sophisticated laboratories. Expert-based evaluations overcome the excesses of employing users, but tend to find facts without problem resolutions. They are also less effective at discovering problems than user-based. Model-based evaluations must undergo validity verification, which may ultimately be highly beneficial but involves a long duration[7].

Rubin defines user-centered evaluations to be *usability testing*. He describes four types:

- "Exploratory" – users try out a prototype or discuss early design.
- "Assessment" – a middle-of-development stage where users test usability of a product's basic functions.
- "Validation" – users test the full product's usability as a whole in preparation for release.
- "Comparison" – at any stage of development users compare designs, versions, competing products, etc.[8]

Nielsen notes that many organizations believe that usability evaluation is too costly and/or too complex to administer. He asserts this not need be the case, and that these beliefs are more likely perceptual than factual. According to Nielsen these perceptions stem from a number of factors and sources, including suspect published reports, misleading models, and simple lack of information[9]. This is interesting in light of Scholz's aforementioned assertion, seemingly endorsed by NIST, that user-centered evaluations are hindered by their expense and the common need to use a lab[7].

Nielsen also describes a related lack of usability research. Researchers want to capture quantitatively significant statistics to justify studying an application's usability. Researchers tend to think it imperative that any tests of an application's usability meet some specific benchmark of confidence, such as would be expected of a drug tested in a clinical trial. Such research may not be attempted in the face of anticipated complexity and an associated financial burden[9].

In response to the above issues, Nielsen developed an evaluation model he termed Discount Usability Engineering[10]. (Nielsen and others in the literature seem not to abbreviate Discount Usability Engineering with an acronym, so I have elected to do the same.) Discount Usability Engineering draws on elements of both user-based and expert-based evaluation, and is appropriate for all types of usability testing[9,10,11]. The idea of Discount Usability Engineering is that in practice a usability evaluation does not need to demonstrate a statistically significant level of confidence; it is not important to find everything that affects usability. What matters is that the usability concerns that are identified deliver the highest cost-benefit ratio[9,10]. Suppose you could only achieve a 25% confidence level that a usability evaluation could improve your product. It is not statistically meaningful – but the chance that you will design the product to be more usable is three times greater than the chance of failing to do so. That might be worth the risk if your costs and effort are low. Nielsen points out that even the percentage of the time you make no improvements, you are unlikely to make the product worse[9].

Discount Usability Engineering is particularly well suited to comparison usability testing, because the primary aim is to figure out which product is most usable rather than to meet a usability threshold. Randomly choosing between two applications gives only even odds of selecting the most usable. If usability evaluation were to reach the aforementioned 3:1 odds, you would have tripled your likelihood of choosing the better application. Via Discount Usability Engineering this increase in confidence level, and more, can often be accomplished with relatively low cost[9].

What enables Discount Usability Engineering to keep usability evaluation at manageable cost and high benefit is that it is designed to use a minimum number of subject users, minimal complexity, and minimal overhead. Nielsen and Landauer determined that the rate of increase in usability improvements discovered per user using Discount Usability Engineering peaks with a small number of testers, sometimes as few as seven. (See **Figure 1.**) Because of this projects can achieve maximum cost-benefit ratio when cost of users is low, resulting in ratios of as high as 1:178 (For every $1 spent, $178 of value is gained). In the same research Nielsen and Landauer learned that after an initial gain additional increases in complexity also do not provide corresponding increases in discovered usability concerns. In other words, while more usability problems might be discovered with greater coverage, the effort to do so would be increasingly less cost-effective than finding the initial problems[12]. Overhead reductions come from methods that avoid expensive videotaping and strategies for reducing evaluation complexity. Nielsen achieves all of these minimalizations by managing Discount Usability Engineering through variations of three usability evaluation techniques: "scenarios", "simplified thinking aloud", and "heuristic evaluation"[10].

Scenarios are users' perspectives on what happens during a transitional process, and as such are a form of user-centered evaluation. There is debate as to whether scenarios are descriptions of paths through the transition versus description of the interactions of the environment with the transition[13]. Nielsen seems to belong to the former camp. He sees scenarios primarily as useful in exploratory or comparison testing, particularly for prototyping[11]. Nielsen uses scenarios in Discount Usability Engineering by creating "extreme" prototypes by vastly limiting product scope. Users need not concern themselves with the applications' overall complexity, thus greatly reducing evaluation cost and administration[9].

Thinking aloud is an old tool often used by psychologists. Persons are asked to verbalize what is going on in their thoughts about an activity while simultaneously engaging in it. The researcher observes the person while recording their comments and behavior on videotape[11]. Simplified thinking aloud is an abridged variation in which the researcher substantially reduces expenses and time by substituting video graphic technology with ordinary note taking[9,10]. Simplified thinking aloud can be either user-centered or expert-based, and appears to be useful in all four types of usability testing.

Heuristic evaluation is the most commonly used expert-based usability evaluation[8]. A group of knowledge realm experts evaluate user interface design against agreed-upon usability rules. Generally each expert conducts an independent evaluation, and their findings are collated after all of them are finished. In Discount Usability Engineering the primary difference from other heuristic evaluation implementations is the use of a considerably smaller rules base. Nielsen states that a typical number is 1000, but that as few as 10 fundamental ones can effect Discount Usability Engineering goals[9].

Analytic Hierarchy Process

As mentioned above, PACM is a computerized automated decision aid that is based upon a decision making model called Analytic Hierarchy Process (AHP). AHP was originally postulated in 1977 by Saaty[14] and explained in much greater detail by him in 1980[15]. Saaty's premise is that when the human mind is confronted with decisions composed of a number of complicated interactions and input, it constructs a "hierarchy … of stratified levels" in order to reach an overall choice[15] [p. *x* (Preface)]. PACM bases its determination of women's future delivery preferences on such a hierarchy (**Figure 2.**)[2].

In AHP decisions are broken down into a number of scored pairwise comparisons between criteria stacked in tiers that Saaty calls levels. Criteria at any given level may or may not have subordinate criteria at the next lower level. There is only one criterion at the top level (the final decision), but there can be any number of subcriteria. Any subcriterion can be subordinated to more than one superior criterion at the next higher level if that level is not the top[15].

Each comparison is constructed between two choices that are not absolute but rather related on a continuum in the manner of a Likert scale. The score on a particular comparison is calculated as a ratio of the extent that the decision maker leans towards one choice to the extent s/he leans towards the other[15]. For example, a user might be given the question, "When you drink beer which of these matter more to you?" The answers might be "Tastes great" and "Less filling". The user would be asked not only which is their preference, but the degree of preference, *i.e.* tasting great is *much more important* than being less filling, or less filling is *slightly more important* than tasting great. The ratio is calculated based on the preference degree. Saaty calculated that the ratios should not exceed 9 and devised a table of rational values to Likert-like labels (**Table 1.**)[15]. For instance, a 5-point scale with the preference on the second label going from left to right would produce a ratio of 5; going from right to left the ratio would be 1/5. On a 9-point scale the same absolute positions (2nd and 8th) would instead produce ratios of 7 and 1/7 respectively. Note that while Saaty's formula does not allow for a ratio greater than 9 or less than 1/9, it does allow for any number of fractional scalar positions[15].

Many decisions are actually more than two-fold, and each decision attribute needs to be compared against every other. In AHP, these attributes are represented by the subcriteria of a particular superior criterion[15]. We can illustrate this using our beer example above;

the decision maker might also want to consider strong hops flavor and low alcohol content. This leads to six different pairwise comparisons:

1) Great taste vs. less filling
2) Great taste vs. strong hops
3) Great taste vs. low alcohol
4) Less filling vs. strong hops
5) Less filling vs. low alcohol
6) Strong hops vs. low alcohol

In the above example, the ratio of 1) might be 6 while the ratio of 2) might be 1/3, and so on.

The essence of AHP is in weighting and normalizing relationships between comparisons, both those adjacent to each other and those of subcriteria to their respective supercriteria. When done properly AHP results in a top-level decision with strength of preference based accurately on the relative contributions of its many input factors.

A plethora of scales, mathematical formulae, and methodologies have been devised to calculate the optimal preference vector. Saaty mentioned four formulae in his early research[15], and later revised that work[16]. Kocaoglu enhanced Saaty's original work by creating the aforementioned Pairwise Comparison Method (PCM)[17]. (Although Kocaoglu described PCM as a different methodology than AHP, he was referring specifically to the work of Saaty, and today PCM is sometimes considered a variant AHP formula.) PCM varies from Saaty's AHP formulae by use of a modified normalization technique called Constant-Sum method. It also replaces Saaty's 9-point scales with ratios between numbers that add up to 100, such that the ratio 3 is represented by 75:25. Other important contributing research includes that by Belton and Gear[18], and Lootsma[19]. By 1994 Triantaphyllou, Lootsma, Pardalos, and Mann were able to evaluate 78 different scales described in research literature[20]. These many flavors of AHP have been used across a number of different industries and disciplines.

For the checkoff version of PACM (Preferences-Assessment Computer Module), Eden *et al* chose to utilize an adaptation of one of Saaty's original formulae[15] to clinical work by Dolan[2,21], and Dolan and Frisina[2,22]. The sliding bar version was an upgrade of PACM as built by Eden, Anderson, Bhupatiraju, and Guise. It uses the ratios from Kocaoglu's Pairwise Comparison Method[5,17], although it performs normalization as described by Dolan and Frisina[22]. In both versions the formulas in PACM use weighted vectors for each group of criteria that are at the same level as each other and subcriteria to the same criterion. These vectors are normalized to add up collectively to 1. Within PACM each of these groups corresponds to a discrete portion of the application with the appropriate comparisons as well as weighted results. The results are displayed as percentages, which are simply the normalized vectors multiplied by 100. Each group of vectors are weighted against each other and normalized again in order to calculate a final score[2]. A detailed explanation of the formulas can be found in **Appendix A**.

Methods and Materials

Research Setting and Subjects

This research was conducted in the Center for Women's Health, an ambulatory clinic at Oregon Health & Science University (OHSU), during the fall of 2004. Three research assistants were employed, although only one research assistant at a time worked with any given subject.

In keeping with the aims of Discount Usability Engineering, a small sample of nine subjects was recruited to compare the two versions of PACM. The results of one subject's experience were dismissed when it became clear that she was unable to understand the research questions. Thus, the data from eight subjects were included in the final evaluation. The recruited subjects were English speaking women at least 18 years of age that had delivered one and only one time, either by VBAC or cesarean, usually within the immediately preceding days or weeks. All subjects were paid for their participation, which was entirely voluntary. None of these subjects were used previously or subsequently in the main study by Eden *et al*. However, Eden *et al* did include all eight subjects in a pilot evaluation of a risk assessment tool inside of PACM.

Six of the eight subjects provided answers to demographic questions. All of the six had completed at least some college in their education, but only two had received a degree (both at the graduate level). All six women used a computer at home, and all were familiar with the Internet. Three of them also used a computer at work.

The Preferences-Assessment Computer Module

As has been discussed above, my research centered around two versions of the PACM application, one in which users responded to pairwise comparisons using radio buttons (the checkoff version) and one in which they responded with a sliding bar. (For examples see **Figures 3.** and **4.)** Both versions of the PACM application display a number of tabs at the top of the screen lined horizontally, each of which corresponds to a separate birth-related topic. Except for the left-most tab, which corresponds to an Introduction, selecting a tab reveals a varying number of additional tabs in a second horizontal row. Most of these additional tabs display a unique page with one or more questions for the user to answer. There are also pages that help the user with definitions but which do not contain questions. The user generally navigates through the pages and tabs sequentially, using a right-pointing arrow at the lower right-hand corner of each page. The user may also navigate backwards using a left-pointing arrow in the lower left-hand corner, or she may jump non-sequentially through the application by clicking on tabs directly.

Some tabs do not consist of pairwise comparisons. Where pairwise comparisons are not displayed, both versions of PACM are identical. Since the purpose of this study was to compare applications, only tabs corresponding to pairwise comparisons were evaluated. In addition, not all comparisons were utilized. This followed the scenario model of Discount Usability Engineering by creating a reduced prototype of the application.

Besides a potentially lowered cost (reflected primarily in reduced working hours of researchers and manageable payments to subjects), the prototype reduction benefited users by requiring less time and effort to run through the application. The tabs that were employed for the study cover comparisons between four birthing experience Criteria, three Side-effects to the mother, and five Delivery factors. These produce six, three, and 15 pairwise comparisons respectively. Every comparison starts with the identical question, "Thinking about your next childbirth, which of these two preferences is more important to you?" Each page differs only in the pair of preferences displayed. The Criteria section also includes an Example tab to demonstrate how to answer pairwise comparisons. From the user perspective it operates identically to the other pairwise comparisons, except that the possible preferences are the meaningless pair "Left criterion" and "Right criterion". Users' responses to the Example are not tabulated in the AHP algorithm.

In the checkoff version of PACM there are 17 radio buttons aligned left to right, with one or the other preference response above each end of the line. The middle button is labeled "Equally important". The buttons two positions to the left and right are labeled "Moderately more important", while those five to the left and right are labeled "Much more important". The far right and left buttons are labeled "Extremely more important". The remaining 10 buttons have no labels. The radio buttons behave according to standard protocol, meaning they are mutually exclusive. Choosing a new button will erase any entry in a previously chosen one.

The sliding bar version of PACM is first viewed as a white horizontal line (the bar), with a small rectangle in the middle (the slider) labeled "50". The slider can be moved back and forth along the bar using the mouse, or by means of holding the mouse cursor over small scrolling arrows at either end of the bar. (These latter arrows are distinct from the navigation arrows in the lower corners of the page.) In contrast to the checkoff version, there are only three labels below the bar: one at each terminus reads "Extremely more important" while the midpoint is marked "Equally important". When a user moves the slider, the bar changes to blue, and the number on the slider increases in single integer increments as the slider approaches the outer ends. At the extreme right and left it is labeled 90. (If the user wants 50 to be her calculated answer, she must move the slider at least one number either direction and then back to 50.) As with the checkoff version, one or the other preference response can be viewed above each end of the line.

Other Materials

Besides the PACM application, two other types of materials were used in this study. Subjects were shown a series of PACM screen shots on paper that were identical to the Example page of the Criteria section (see **Appendix B**). Each screen shot was created in both a checkoff and sliding bar copy; for any given pair of screen shots the potential answer was placed in a specific position that was computationally similar (per AHP) in both versions. These screen shots were used as props while asking users a series of questions about their understanding of the application.

The screen shot questions were designed to assess whether or not use of the application was relatively intuitive. These questions zeroed in on users' understanding of the use of the application, to determine if they knew how the application was supposed to work and what the placements of answers to the pairwise comparisons represented. Users were also asked a series of open-ended questions concerning their opinions of facets of each just-used version. These ended with a lone Likert scale based question to assess the ease or difficulty of their experience. Finally, after both versions were completed users were asked which version they preferred, followed by demographic questions concerning their levels of education and computer use. For a full list of the questions and screen shots, see **Appendix B**.

Methodology

This evaluation explored both the scenario and simplified thinking aloud techniques of Discount Usability Engineering. Both were fundamental in the design of the evaluation, as will be described below. However, because of a lack of usability experts it was not possible to conduct a heuristic evaluation.

This evaluation was structured largely as qualitative research. The first phase of the project was a combined assessment of the user questionnaire and training exercise for the research assistants. Three women associated with the Biomedical Informatics department at OHSU and one outside the university were asked to follow through the same routine as would later be used by subjects. This routine began with each user being given an explanation of the purpose of the application and of the exercise. It was also explained to them that their answers would not be recorded nor used for calculating their overall birthing choice. One or the other version of PACM was chosen by a simple coin flip, opened, and set up such that the user would begin with the Criteria tab and the Example sub-tab. Each user was asked to use PACM as they perceived was appropriate but not given instructions. As suggested as part of Discount Usability Engineering, users were encouraged to "think aloud" as they went through the application. They were asked to comment on any aspect that caused them problems or confusion, or that they liked. Instruction was given to follow the application in sequence (per the right-pointing arrows) through all 24 pairwise comparisons and then halt. Once they were finished I showed them each of the screen shots for the specific version and asked all the understanding and opinion questions. When all the questions were asked, the entire exercise from the Example page to the last question was repeated for the opposite version of PACM. Each user was then asked their choice of screen, and finally asked to comment on the questions, screen shots, and any other aspect of the exercise itself.

As a result of this preliminary phase a small number of items were changed in the questions. More significantly, the Example page was itself altered within PACM and on the associated screen shots. When these trial users worked PACM, the Example page had the same pairwise comparison as the first Criteria page. As a result of the preliminary exercise it was realized that if users suggested a particular response was more meaningful than any other it would be impossible to determine if it was because it was their preferred choice or because they did not understand the application. Thus, the Example page and all screen shots were remade using the completely neutral "Left Criterion" and "Right Criterion" responses.

Once the research materials were designed and created, the evaluation was conducted with real subjects, each of whom was given identical tasks with both of the two application versions. The version of the screen used first was selected randomly beforehand by an off-site computerized randomizer. As with the four trial users, subjects were given an explanation of the study, but not given instructions, and they were informed that their responses would not be recorded nor used for a final calculation. All subjects were observed answering the same questions in each version of the application while being encouraged to think aloud, during which the research assistants took field notes. Each subject used one version of the application, was then questioned about her understanding and opinions of it, and then repeated the sequence with the opposite version. In the end each subject was asked which application she preferred and her reasons for her preference, followed by a couple of demographic questions. Once a subject completed all questions, she was offered time to use the application in its normal function to assess her own birthing preferences. This latter opportunity was offered as a voluntary perk outside of the research, and the users' responses were not observed to insure privacy.

Because of the small sample size, and because it is not the aim of Discount Usability Engineering, no quantitative statistical analysis was performed on the data. However, I did recalculate representative PACM scores using the values that users commonly perceived they were entering, and compared them to the scores from the values actually entered by the AHP algorithm. While the results of these comparisons cannot be considered statistically significant, in the spirit of Discount Usability Engineering they revealed useful improvements to recommend to Eden *et al* for minimal cost.

Results

Preferred Version

There was a slight overall preference by users for the checkoff version. Four of the eight users preferred the checkoff version, while three preferred the sliding bar. The eighth subject was undecided. These results were unrelated to the order in which they were exposed to the two versions. Three subjects preferred the version to which they were first exposed, while four preferred the second version they tried out.

However, none of these results are significant or even particularly revealing with such a small sample size. More importantly, the subjects' preferred versions were not necessarily reflective of their understanding of either version nor were they evidence of inherent usability.

Field Observations

During observations it was noted that subjects generally responded to comparisons using those radio buttons that had labels, or slider values divisible by five. Because users were encouraged to think aloud, some subjects expressed that they felt there were too many settings. Several expressed surprise that the slider had no value greater than 90; they suggested that an end value of 100 made much more sense to them. Likewise a desire to start the slider at 0 or 1 instead of 50 was mentioned. At least one user tried in futility to move the slider past 90 by holding down a scrolling arrow.

Both subjects and preliminary phase users expressed interpretations of the sliding bar values as percents, which was given as the reason for expecting the high value to be 100. Some of them described a setting of 75 as halfway between the middle and the end, even though it was not physically at that point. Yet they correctly understood 75 as meaning that one preference was 3 times as important as the other preference.

One of the preliminary phase users had trouble understanding that the buttons were mutually exclusive, or that a comparison allowed no more than one preference setting. She attempted to mark one button each on either side of the middle. Later she expressed that she had been trying to choose a relative degree of preference to each choice in the comparison. None of the actual subjects displayed or expressed this confusion, but one woman always first marked one of the buttons labeled "Extremely more important" and then reset her answer on one of the other labeled buttons on the same side.

At least three subjects were confused by the choices themselves in the Example page (and on the screen shots). One of these women was not familiar with the word "Criterion" (which was used generically for the usability test only), and all three expressed difficulty choosing because the choices had no meaning to them.

One subject felt that it was not helpful to read the question before making a choice, particularly since the question was always the same. This subject also brought to attention her belief that the choice "Avoid death or disability to my baby" was not meaningful, because it was likely to be the preferred choice in any comparison.

The sliding bar version of PACM did not register the value 50 unless users moved the slider to another value and back to 50. This was necessary to make certain the application could differentiate a choice of 50 from a failure to answer the question. Yet some found this frustrating and felt it should be possible to register 50 by default.

Understanding of PACM

All eight subjects generally understood that responses to the left of the middle were oriented towards the answer on the left end while responses to the right were oriented towards the right-hand answer. Most subjects also expressed that the middle represented both choices being equally important, although one suggested it simply meant the user could not make up her mind. (The difference between these is subtle, but should not be overlooked – a user that does not make up her mind may only have failed to complete the data entry rather than revealed her actual choice.) Finally, subjects universally understood that an answer placed in an end position meant that the choice at that end was preferred as much over the other criterion as the application allowed.

For the most part subjects recognized that responses further away from the middle indicated a greater degree of preference for a particular criterion than responses closer to the middle. However, there was not a consensus concerning the meaning of specific placements in between the middle and the ends. For instance, the first screen shot of the checkoff version showed a user's preference placed four buttons to the right of the middle. This button was the right most of the two positions in between buttons labeled "Moderately more important" and "Much more important" on the right-hand side (refer to **Figure 3.**). When asked, "Can you please tell me what this selection means?" only two subjects responded in a manner suggesting detailed understanding, with one stating that the user's choice was moderate to much more important and the other suggesting that the choice was "closer" to much more important. Two others indicated that the user's preference was simply much more important (*i.e.* as if the position was labeled), and three responded only that it was more important. The remaining subject said that the user was undecided, although it was not clear whether she meant undecided about the two criteria or just undecided about the two nearby labels.

Questions about the sliding bar unveiled lack of consensus even in the meaning of numeric values. The sliding bar screen shot that most closely approximated the intent of the aforementioned checkoff picture displayed the slider with the value 69. Four subjects stated this showed that the user preferred the right criterion or it was more important than the left, but did not qualify their interpretation any more deeply. Another subject said the user considered the right "somewhat" more important and one other said it meant "moderately" more important. The remaining two subjects referred to the actual value displayed in the slider. One explained the number to mean that the right criterion was 19 points greater than the left, while the other commented that the right was "69% more weighted that the left; the left would be 31%".

For the checkoff version subjects were shown two screen shots that had the left side "Moderately more important" button checked in one and the left-side "Much more important button" checked in the other. When asked how these various displayed answers differed, most subjects understood that the response "Much…" signified greater preference over "Moderately…", but they described the separation in different degrees: "much more", "more", "slightly more". One subject said only that the two pictures represented a user changing her answer. Similarly, for the sliding bar version subjects were shown a pair of screen shots with the values on the left at 60 and 75 respectively. Some of the subjects described the difference with Likert-like phrases, but four of them used purely arithmetic terms, *i.e.* 15 points or 15 percent.

The subjects were also shown two checkoff screen shots in which the first was marked "Moderately more important" but the second shot displayed the preference in the unlabeled button immediately adjacent to the left of "Moderately more important". This time the users answered with more uniform intent, with all but one stating the difference in some manner representing a minimal amount (most used the word "slight"). The lone outlier again thought only that the user changed her answer. However, two of those that saw a slight difference were not comfortable with the settings. One stated she might move her response from one radio button to the adjacent one "just to shake things up" so as to avoid all responses being the same. The other responded with dismay that there were unlabeled buttons at all and said she would not herself use them.

The related pair of sliding bar pictures displayed settings at 60 and 62. Subjects found these much more difficult to differentiate. Three subjects expressed that the closely placed responses in these pairs might not represent different intentions. One of these thought that adjacent slider values might be caused by an accidental movement of the mouse.

Opinions of PACM

Subjects found it relatively easy to determine where to place their answers in both versions, although not always to the same degree. Seven of the subjects described the checkoff version in phrases such as "very easy", "somewhat easy", "wasn't too hard", and "4 on a 1-10 scale with 10 = hardest". The one dissenter described it as "pretty difficult". Along the same lines six subjects described the sliding bar using terms like "easy" or "very easy". One of the remaining two felt it was hard, and the other simply said she did not like sliding. However, when asked how easy it was to view their answers, the sentiments for the sliding bar were comparable to the prior question (in fact, most said "very easy"), but four subjects were not completely satisfied with the checkoff version. These subjects stated that the size of the buttons was too small and two expressed that they were not spaced adequately.

Five subjects were happy with the number of positions to place an answer on the checkoff version. Two believed that the number should be reduced by four positions, and the other subject just responded that she preferred the sliding bar. Five subjects were also satisfied with the number of values on the sliding bar, but in contrast to the checkoff version the dissenters felt the problem was not enough values. All three of them wanted a scale that reached 100, and one of them felt it should start at 1.

Most of the subjects considered the checkoff version appealing as is. However, one felt that it needed a better font, and another subject called it "boring" without offering suggested improvements. Only three subjects considered the sliding bar version appealing. Some complaints: "The numbers don't tell you anything", and, "[It] was a little confusing because you had to click on 50 to make that choice before you could move on." Suggested improvements for the sliding bar included making the box stand out more, using more or brighter color, a thicker bar, and a larger font for the question.

Although a few subjects felt there were too many buttons in the checkoff version, none wanted additional labels under the unlabeled buttons. All the subjects were satisfied with the labeling of the buttons as they were. A couple thought that the application would be usable without labels, but none preferred that solution. By contrast, only one subject saw a need to add text labels under the sliding bar. Even though they were being asked about labels under the bar, four subjects again expressed dissatisfaction with limiting the values at 90 instead of 100, with one suggesting that the middle should be value 0 (not the subject that had earlier suggested 1). One subject thought that removing the numbers altogether would improve use of the slider.

When subjects were asked if they could think of a way to mark answers that would be better either than the version they had just seen (if it was the first) or both versions (if they had used both), half the subjects had no suggestions. The remaining half all suggested checkboxes, which is interesting in that such a solution would allow multiple responses to the same comparison.

In the final question, subjects were asked to rate the difficulty of its use on a Likert scale, as "very difficult", "somewhat difficult", "neither difficult nor easy", "somewhat easy", or "very easy". No subject rated either version as either very difficult or somewhat difficult. The results of their ratings are in **Table 2.** (One subject did not rate the checkoff version.)

Discussion

I would like to think, as Nielsen and Landauer would predict[13], that with only eight subjects this research covered the majority of usability issues that could ever be found. Certainly those that have been discovered were done so with low cost. Truly, though, many of the matters brought to light were purely cosmetic matters. Some good suggestions were made concerning fonts, button sizes, and amount of color, and no subject argued against them. Yet it did not appear that any user was impeded or heavily impacted by these considerations. In several cases a majority of the users liked the status quo with both versions. They were happy with the labels; they found the checkoff version appealing; they thought it was easy to figure out where to place their answers. It could be argued that from an aesthetic point of view both versions of the application were favorably received as they were presented. If this was all there was to the evaluation, even at a discount it is hard to see the cost-effectiveness.

However, usability evaluation is not a marketing research tool; it is a facet of design and engineering. While satisfaction is one of Nielsen's five components of usability[6], it does not stand alone. What is notable about the subjects' results is the extent to which they misunderstood the meaning of their answers. On both scales subjects had a hard time understanding the difference between close values, and they avoided unlabeled buttons or values not divisible by 5. Some suggested that there were too many buttons on the checkoff version, not from an aesthetic point of view but from an operational one – they saw them as unchoosable.

It appears that most subjects understood that the sliding bar numbers represented percentages, but not that they interpreted the percentages the same way. Those who saw 75 as representing 75% for one answer versus 25% for the other appeared to grasp the concept as intended. Yet at least two subjects believed that there should be 100 values on each side. This would seem to indicate a desire that 75 represent 175 out of 200, which is 87.5%.

Also notable is the fact that subjects avoided numeric values not divisible by 5, and that several could not differentiate values only 2 apart. As explained in **Appendix A**, ratios of 2, 5, 6, 7, and 8 do not fall upon divisions of 5; therefore users do not seem likely to choose a preference because it is 2 or 5 times as strong as the other. (In fact, the precise ratios are not even integers, so they would be unable to choose these exact ratios even if desired.) Finally, several subjects expressed frustration that the high score of 90 did not represent their full preference level of 100% (which cannot be expressed as a mathematical ratio because it would require division by zero).

With the checkoff version there was no evidence in the study that users associated radio button settings with any particular numeric value. Therefore, there is no reason to assume that their intent was different than the measure. However, the subjects in this study did not necessarily understand the labels under the buttons, as they described the same position with differing degrees of importance. It should be noted that this occurred with a position that some said they would never choose and that most ignored during observation. This suggests not so much that users choose incorrect representations of their preferences as that they are not as discriminating in their choices as the formula allows.

These matters signified major implications to the efficacy of PACM. The premise of AHP is that overall choice is complex, so decision attributes must be carefully dissected, stacked, and weighted for proper decision support[15]. Yet what if the women can't quantify the preferences as precisely as the program requires? How much might this affect the outcome of the top-level decision?

It is hard to say what the cumulative effect would be for all pairwise comparisons for any individual user. Even in the worst case for the sliding bar, the attribute weights are in the same relative order. Ultimately it may just be that overall decisions tend to be representative of the users' intentions from a binary A versus B perspective, but the strength of a decision may be deemed less or greater than it should be.

Finally I note the curious suggestion by half of the subjects that checkboxes would be a better solution than either radio buttons or a sliding bar. None of the subjects explained how they would anticipate the mathematics to work, but one would assume it would have to require some sort of adding or averaging of values. Apparently at least one of the preliminary phase users had exactly that idea in mind as she tried to enter two values for each checkoff preference. The subjects, of course, had no knowledge of the formulas used in PACM. Therefore they cannot have appreciated the possibility that checkboxes have no merit as a workable decision support methodology. Yet from a usability perspective, they raise the consideration that AHP isn't necessarily satisfactory.

Unfortunately, there is not much precedent from which to draw conclusions. It is perhaps a reflection of the research world reluctance described by Nielsen[9] that I could locate only the scantiest reference in literature to usability evaluation for AHP based applications and none for PCM based. Shepperd and Cartwright conducted a limited usability evaluation of a project management tool based in part on AHP, and they found that student users had some trouble making pairwise decisions through a user interface[23]. Unfortunately they did not describe the design of the interface, so it is unclear to what extent it bears relevance to my research. Asahi, Turo, and Schneiderman briefly mention usability testing of an AHP tool in their research, but give no detail[24]. Canfora and Troiano stated as recently as 2003 that they were unable to find any published research describing usability of software products built upon AHP or other similar decision making techniques[25]. (Interestingly, some researchers have used AHP as a usability engineering tool in applications that do not themselves incorporate AHP, including Sikorski[26], Baek, Liebowitz and Lewis[27], and Levin[28].)

I found no literature at all describing usability evaluation for AHP based applications in the clinical world. In fact, there are very few published examples of Discount Usability Engineering on clinical decision support applications even built without AHP. One notable exception concerns usability evaluation by Carroll, Marsden, Soden, Naylor, New, and Dornan of a tool to advise patients and physicians of cardiovascular risk. Carroll *et al* were able to demonstrate the efficacy of Discount Usability Engineering with a smaller clinical decision support application that, like PACM, allowed direct manipulation by patients as users.[29] Nonetheless, it appeared I had traveled into nearly virgin territory. Clearly, further research needs to be done in this area

Perhaps the overriding suggestion of the above results is that PACM suffers from a lack of inherent learnability. The correct use is evidently not intuitive. However, in this study the subjects were given no instructions. It is arguable that simply explaining what the potential responses truly represented might have yielded considerably different results. This is in line with Patel and Kaufman, who describe learnability as separate from usability within medical informatics products, because of the greater sophistication of the medical world. They point out the need for training and support for such products to be usable[30].

Sadly usability evaluation remains a surprisingly overlooked feature of computer applications development. Perhaps the reader has dealt with software or Internet applications that require an irritating series of back and forth navigations, data that is difficult to locate, or menu selections that are overlapping and ambiguous. These frustrations are well-known examples of computer products missing the mark on usability.

Rubin gives several cogent reasons for the common failure to achieve usability:

- Product engineering pays much more attention to hardware and information technologies than to their users.
- The user community has evolved without concomitant recognition of this phenomenon by the developers.
- Usability is not intuitive, yet engineering organizations often think it is.
- User interface and documentation teams communicate poorly with systems teams.
- Engineers are taught technical expertise while product features require design expertise[8].

When one considers that AHP is a technical model rather than a user interface system, it is perhaps then not surprising that a usability evaluation of PACM produced usability issues surrounding AHP that needed to be addressed. On the other hand, to the credit of Eden *et al*[2] and to Eden, Anderson, Bhupatiraju, and Guise[5], in many other ways PACM was already a reasonably usable application.

It should be noted that aspects of the research itself confused some users, and this should not be overlooked. It may be argued that if users had dealt with real preferences in the Example, or more importantly, in the screen shots, instead of the irrelevant "Left criterion" and "Right criterion", that they might have shown more understanding of the preference choices available. My guess is that this was really more an exposure of users' difficulty making choices that are of no value to them rather than evidence of an inability to understand the choice mechanism. However, future researchers may want to consider an improved method for gauging users' understanding of applications employing pairwise comparisons.

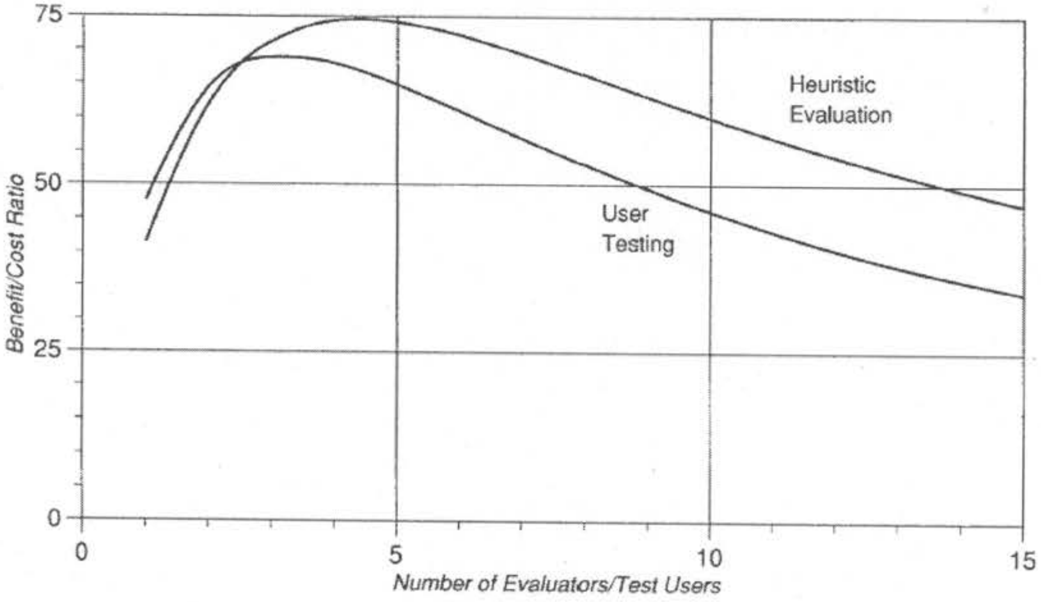
Summary and Conclusions

From the evidence collected in this research I concluded that the checkoff version of PACM was more usable than the sliding bar version. This is not because four users preferred the checkoff version versus three that preferred the sliding bar. I concluded this instead because I believed that the research indicated the checkoff version caused less confusion and provided more consistent usage. Not all persons using the sliding bar interpreted its values identically, except at the middle and ends. Many expressed frustration with the upper boundaries, or stated that they found no meaning in the numeric values. Finally, a greater number of users found it unappealing as compared with the checkoff version.

Nonetheless, the usability of the checkoff version was open to improvement. These were my recommendations:

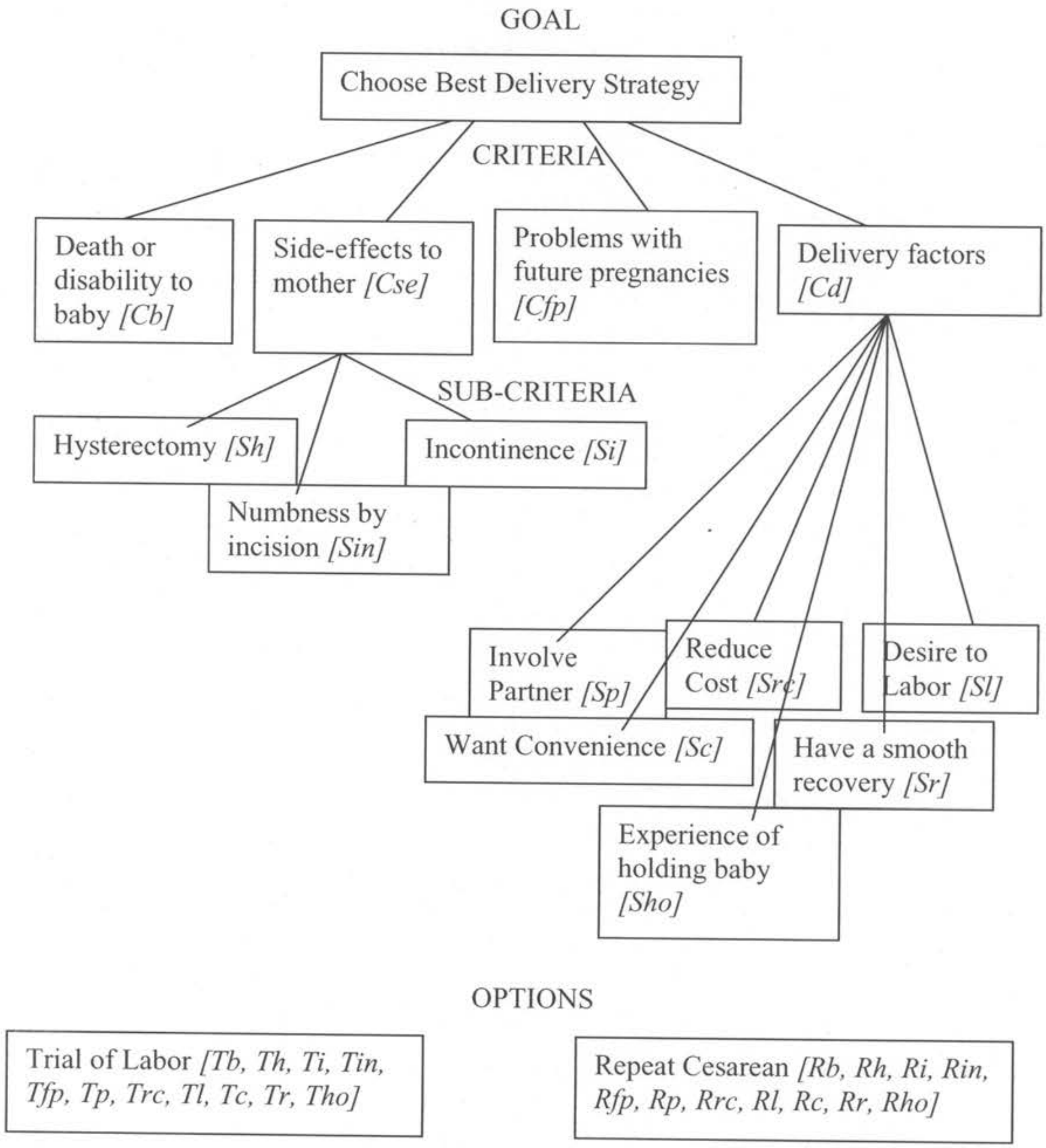
- Removal of the unlabeled buttons.
- Larger size of the buttons.
- Greater spacing of the buttons (implicit with the above removal).
- Color improvements.

AHP is an oft-used decision support methodology that has been deployed in many applications with almost no investigation of its affect on usability. In this study I believe I demonstrated that AHP is not always an inherently usable system, and that careful examination and instruction may be needed for decision support tools built with AHP. At the same time I was able to provide a low-cost yet cost-effective usability evaluation of an AHP based tool using Discount Usability Engineering, which yielded meaningful suggestions for usability improvements. Discount Usability Engineering may be an effective system for future evaluations of AHP based tools.



Ratio between benefits and costs for using various numbers of heuristic evaluators and test users to find usability problems in a medium-large software project, as calculated using the various assumptions listed in the text.

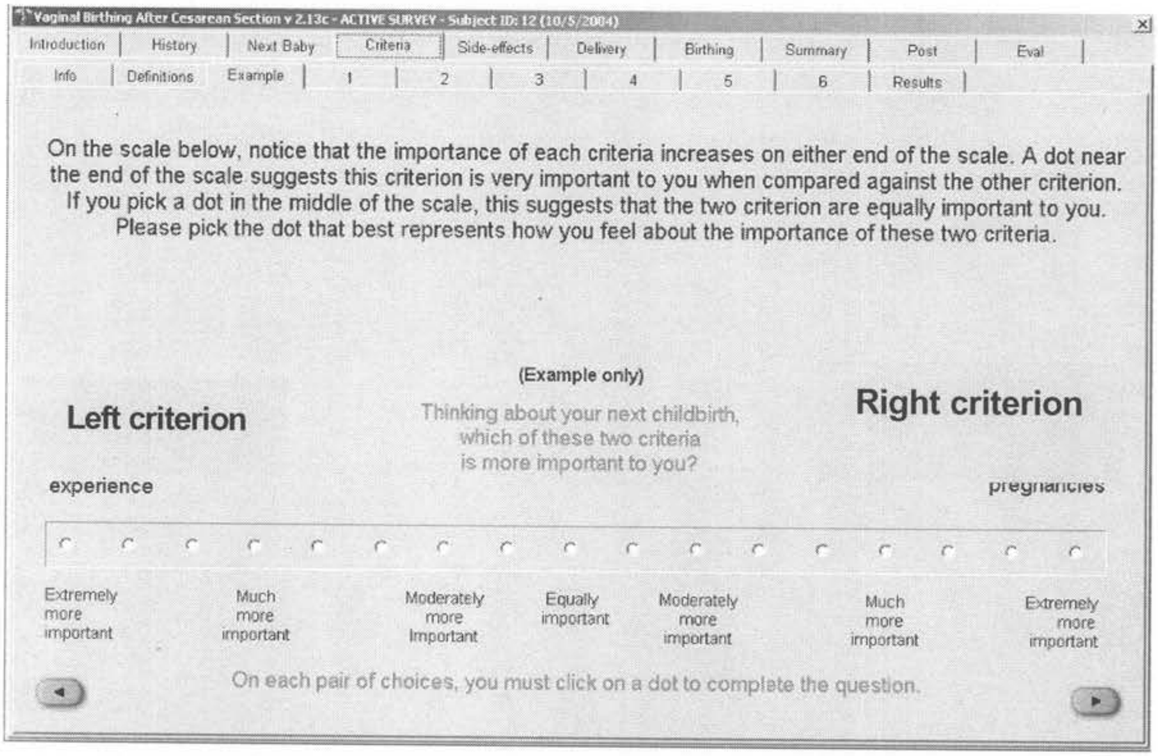
Figure 1. Nielsen and Landauer comparison of number of users to cost-benefit ratio[9]


GOAL
Choose Best Delivery Strategy
CRITERIA
Death or disability to baby [Cb]
Side-effects to mother [Cse]
Problems with future pregnancies [Cfp]
Delivery factors [Cd]
SUB-CRITERIA
Hysterectomy [Sh]
Incontinence [Si]
Numbness by incision [Sin]
Involve Partner [Sp]
Reduce Cost [Src]
Desire to Labor [Sl]
Want Convenience [Sc]
Have a smooth recovery [Sr]
Experience of holding baby [Sho]
OPTIONS
Trial of Labor [Tb, Th, Ti, Tin, Tfp, Tp, Trc, Tl, Tc, Tr, Tho]
Repeat Cesarean [Rb, Rh, Ri, Rin, Rfp, Rp, Rrc, Rl, Rc, Rr, Rho]

[Note: letters in brackets are weights computed by computer program as patient enters her data.]

The patient will have two waves of comparisons. In the first wave, she will be given definitions for each criteria and sub-criteria. She will make comparisons among them for importance to the delivery decision. Next the patient will be given the probabilities (for the general population) related to each criteria and sub-criteria stratified by each delivery option. The patient will then make comparisons between the two delivery options related to risk, side-effects, and desired delivery experience.

Figure 2. PACM analytic hierarchy of women's birth preferences.[2]



Vaginal Birthing After Cesarean Section v 2.13c - ACTIVE SURVEY - Subject ID: 12 (10/5/2004)
Introduction | History | Next Baby | Criteria | Side-effects | Delivery | Birthing | Summary | Post | Eval
Info | Definitions | Example | 1 | 2 | 3 | 4 | 5 | 6 | Results

On the scale below, notice that the importance of each criteria increases on either end of the scale. A dot near the end of the scale suggests this criterion is very important to you when compared against the other criterion. If you pick a dot in the middle of the scale, this suggests that the two criterion are equally important to you. Please pick the dot that best represents how you feel about the importance of these two criteria.

(Example only)

Left criterion

Thinking about your next childbirth, which of these two criteria is more important to you?

Right criterion

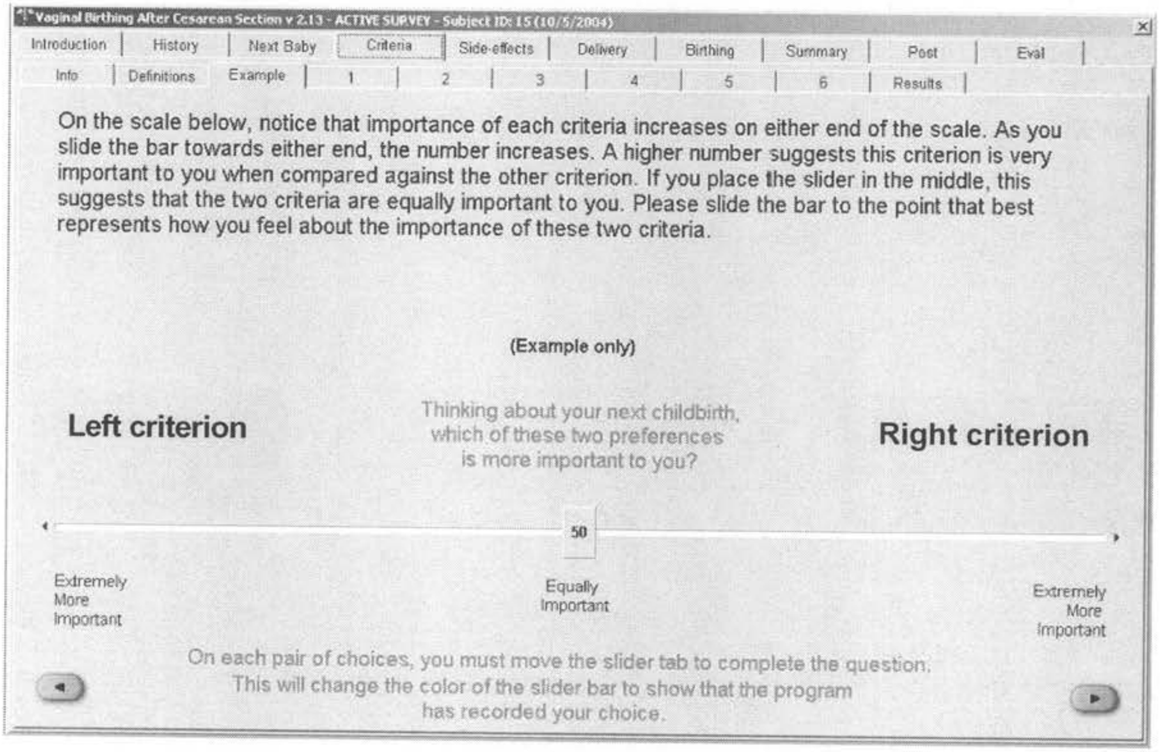
experience

pregnancies

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

On each pair of choices, you must click on a dot to complete the question.

Figure 3. Example of PACM checkoff version using radio buttons



Vaginal Birthing After Cesarean Section v 2.13 - ACTIVE SURVEY - Subject ID: 15 (10/5/2004)
Introduction | History | Next Baby | Criteria | Side-effects | Delivery | Birthing | Summary | Post | Eval
Info | Definitions | Example | 1 | 2 | 3 | 4 | 5 | 6 | Results

On the scale below, notice that importance of each criteria increases on either end of the scale. As you slide the bar towards either end, the number increases. A higher number suggests this criterion is very important to you when compared against the other criterion. If you place the slider in the middle, this suggests that the two criteria are equally important to you. Please slide the bar to the point that best represents how you feel about the importance of these two criteria.

(Example only)

Left criterion

Thinking about your next childbirth, which of these two preferences is more important to you?

Right criterion

On each pair of choices, you must move the slider tab to complete the question. This will change the color of the slider bar to show that the program has recorded your choice.

Figure 4. Example of PACM version using sliding bar

Intensity of importance	Definition	Explanation
1	Equal importance	Two activities contribute equally to the objective
3	Weak importance of one over the other	Experience and judgment slightly favor one activity over another
5	Essential or strong importance	Experience and judgment strongly favor one activity over another
7	Very strong or demonstrated importance	An activity is favored very strongly over another; its dominance demonstrated in practice
9	Absolute importance	The evidence favoring one activity over another is of the highest possible order of affirmation
2,4,6,8	Intermediate values between adjacent scales	When compromise is needed
Reciprocals of above zero	If activity *I* has one of the above nonzero numbers assigned to it when compared with activity *j*, then *j* has the reciprocal value when compared with *I*	A reasonable assumption
Rationals	Ratios arising from the scale	If consistency were to be forced by obtaining *n* numerical values to span the matrix

Table 1. Saaty's scale of AHP references.[15]

	Very difficult	Somewhat difficult	Neither difficult nor easy	Somewhat easy	Very easy
Checkoff	**0**	**0**	**2**	**1**	**4**
Sliding Bar	**0**	**0**	**1**	**3**	**4**

Table 2. Subjects' ratings of overall ease of use for both PACM versions

Appendix A.

The AHP Formula Used in the Preferences-Assessment Computer Module

To understand the formula used by PACM the reader should refer to Figure 2. for the source of specific decision weights. In the figure, note that all weights are shown in brackets, beginning with a single upper case letter followed by one or two lower case letters. All weights beginning with the same upper case letter are between attributes of the same decision criterion. Each of these attributes is compared to all the other attributes of its criterion to arrive at its own weight using a 9 point scale in which ratios between attributes may range from 1 to 9 from the middle to the left end and 1 to 1/9 from the middle to the right end. (Note that in any given comparison, the ratio represents the preference to avoid a specific outcome, not to attain it.)[2]

In the checkoff version of PACM there are 17 buttons, with each representing an individual AHP ratio. The middle button is labeled "Equally important" and represents 1. As the buttons progress to the left they stand for 2, 3, 4, 5, 6, 7, 8, and 9 in sequence. Only buttons 3, 6, and 9 are labeled, as "Moderately more important", "Much more important", and "Extremely more important" respectively. Buttons to the right of "Equally important" fall in a reverse pattern of 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8, and 1/9 with labels at 1/3, 1/6, and 1/9.

The sliding bar version of PACM is much more granular. It is labeled only at the middle and endpoints, but it is marked at all entry points by numeric values. . The middle position is valued at 50. As the user slides the bar to either the left or right the value increases in integer increments up to 90 at the far ends. In contrast to the checkoff version the sliding bar uses PCM ratios. At any given value p on the left side the PCM ratio is represented by $p / (100 - p)$. The corresponding value on the right is represented by 1 / PCM, which is equivalent to $(100 - p) / p$. Thus a left side position of 75 represents a PCM ratio of 75 / (100 – 75) or 3, while the same number on the right works out to 1 / 3. As with the checkoff version the midpoint equals a ratio of 1, and bar value 90 represents ratio 9 or 1/9. However, the ratios in between are logarithmic. The approximate values corresponding to PCM ratios 2 through 8 are 67, 75, 80, 83, 86, 88, and 89 respectively.

The first step is normalization. To achieve this, ratios first need to be placed in a matrix of decision attributes. Each attribute is represented both in a horizontal row and a vertical column, and as a result each pair of attributes intersects twice within the matrix. The ratios between two attributes are indicated at their intersections, with one ratio an integer and the other in the form 1/n. Where attributes intersect themselves, the ratio is always 1. Next, each column should be summed. Then each ratio within a column must be divided by its column's sum. The results of these calculations are the normalized ratios[22].

Once a normalized matrix is completed, the next step is weighting. Each normalized row in the matrix should be summed. Each row sum is then divided by the number of attributes in a row. The resulting calculations are the weights for each row attribute[22].

Let us illustrate this process with an example from PACM. Hysterectomy [*Sh*], Numbness by Incision [*Sin*], and Incontinence [*Si*] are all attributes of one criterion, Side-effects. Suppose that a user places her response for the importance of avoiding a Hysterectomy versus avoiding Numbness by Incision in the radio button marked "Much more important" to the left of "Equally Important". The user then checks the unlabeled left-hand button that is one position to the left of "Moderately more important" for avoiding Hysterectomy versus avoiding Incontinence. Finally she checks the right-hand button labeled "Moderately more important" for avoiding Numbness by Incision versus avoiding Incontinence (selecting on the right means she prefers to avoid Incontinence). Since the middle position corresponds to 1, then the answer placements are in positions 6, 4, and 1/3 respectively. This results in the following matrix of ratios:

	Hysterectomy	Numbness by Incision	Incontinence
Hysterectomy [*Sh*]	1	6	4
Numbness by Incision [*Sin*]	1/6	1	1/3
Incontinence [*Si*]	1/4	3	1

To arrive at each weight, we sum all the ratios in each column, and divide each ratio in a column by that column's sum. The above results from Side-effects will produce the following adjusted matrix:

	Hysterectomy	Numbness by Incision	Incontinence
Hysterectomy [*Sh*]	0.71	0.60	0.75
Numbness by Incision [*Sin*]	0.12	0.10	0.06
Incontinence [*Si*]	0.18	0.30	0.19

Finally, we add all the results in a particular row and divide by the number of items in a row to get that row attribute's weight. Continuing with our Side-effects example we find:

$$Sh = (0.71 + 0.60 + 0.75) / 3 = 0.69$$

$$Sin = (0.12 + 0.10 + 0.06) / 3 = 0.09$$

$$Si = (0.18 + 0.30 + 0.19) / 3 = 0.22$$

Note that the total of all three attributes approximates 1.0. The formula inherently normalizes all values for the same criterion.

Once all the attributes of each criterion are weighted, scores are determined for each of the two main delivery options, VBAC and repeat cesarean, as per an AHP method known as Distributive[2] and utilized by Dolan and Frisina[22]. This results in the following final equations:

VBAC = $(Tb*Cb) + (Th*Sh*Cse) + (Ti*Si*Cse) + (Tin*Sin*Cse) + (Tfp*Cfp) + (Tp*Sp*Cd) + (Trc*Src*Cd) + (Tl*Sl*Cd) + (Tc*Sc*Cd) + (Tr*Sr*Cd) + (Tho*Sho*Cd)$

Repeat Cesarean = $(Rb*Cb) + (Rh*Sh*Cse) + (Ri*Si*Cse) + (Rin*Sin*Cse) + (Rfp*Cfp) + (Rp*Sp*Cd) + (Rrc*Src*Cd) + (Rl*Sl*Cd) + (Rc*Sc*Cd) + (Rr*Sr*Cd) + (Rho*Sho*Cd)$

PACM also watches for inconsistency. For instance, if in the Side-effects example the user had placed her answer to the right on all three ratios, then it would have produced the illogical combination of relationships A>B, B>C, and C>A. In some cases the answers may come out wrong in a more subtle manner: A>B, B>C, and A>C, but the difference between A and C is less than the sum of the differences between A and B and between B and C. PACM catches these types of problems and lets the user know that calculations that reach these conclusions are inconsistent. It advises that the user should reconsider some of her pairwise comparisons[2].

Appendix B.

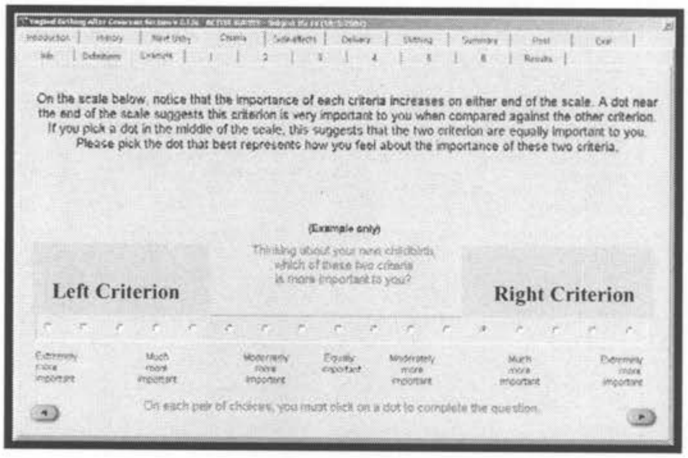
Survey Questions and Associated Screen Shots

CHECKOFF SCALE

The following questions with pictures are to determine the subject's understanding the Checkoff Scale:

Instruct her to explore its use until she's satisfied that she understands how it works. At that point she may continue on her own. She should answer all application decision questions in the Criteria and Side Effects sections by herself. When she reaches the first Info page of the Delivery section, have her stop. At this point use the questions below.

1. Show picture with "Left Criterion" as choice on left end and "Right Criterion" on the right, setting two positions to the right of "Moderately more important"


Introduction | History | Next Baby | Criteria | Side-effects | Delivery | Stitching | Summary | Post | Exit |
Info | Definitions | Example | 1 | 2 | 3 | 4 | 5 | 6 | Results |

On the scale below, notice that the importance of each criteria increases on either end of the scale. A dot near the end of the scale suggests this criterion is very important to you when compared against the other criterion. If you pick a dot in the middle of the scale, this suggests that the two criterion are equally important to you. Please pick the dot that best represents how you feel about the importance of these two criteria.

(Example only)

Thinking about your next childbirth, which of these two criteria is more important to you?

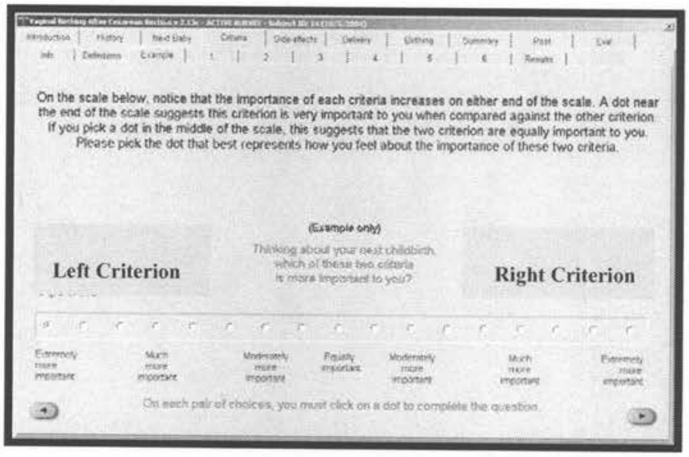
Left Criterion **Right Criterion**

○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○

On each pair of choices, you must click on a dot to complete the question.

Suppose you saw another person's answers, and she placed her selection as on this screen. Can you please tell me what this selection means?

2. Show picture with "Left Criterion" as choice on left end and "Right Criterion" on the right, setting at far left


Introduction | History | Next Baby | Criteria | Side-effects | Delivery | Stitching | Summary | Post | Exit |
Info | Definitions | Example | 1 | 2 | 3 | 4 | 5 | 6 | Results |

On the scale below, notice that the importance of each criteria increases on either end of the scale. A dot near the end of the scale suggests this criterion is very important to you when compared against the other criterion. If you pick a dot in the middle of the scale, this suggests that the two criterion are equally important to you. Please pick the dot that best represents how you feel about the importance of these two criteria.

(Example only)

Thinking about your next childbirth, which of these two criteria is more important to you?

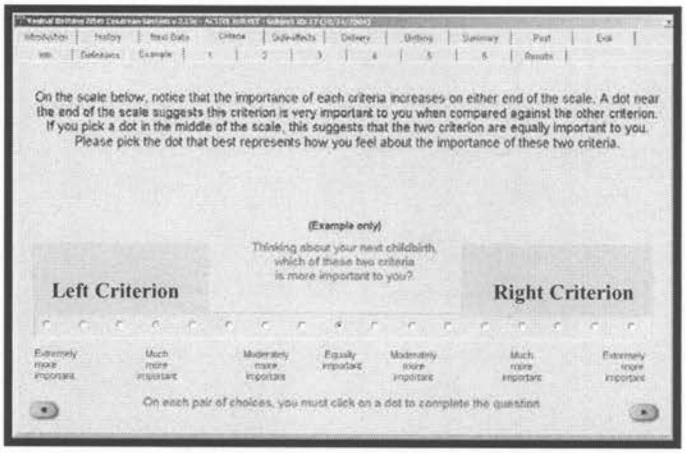
Left Criterion **Right Criterion**

● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

On each pair of choices, you must click on a dot to complete the question.

Now can you please tell me what this selection means?

3. Show picture with "Left Criterion" as choice on left end and "Right Criterion" on the right, setting on "Equally important"


Introduction | History | Next Data | Criteria | Side-effects | Delivery | Birthing | Summary | Past | Exit
Intro | Definitions | Example | 1 | 2 | 3 | 4 | 5 | 6 | Results

On the scale below, notice that the importance of each criteria increases on either end of the scale. A dot near the end of the scale suggests this criterion is very important to you when compared against the other criterion. If you pick a dot in the middle of the scale, this suggests that the two criterion are equally important to you. Please pick the dot that best represents how you feel about the importance of these two criteria.

(Example only)

Thinking about your next childbirth, which of these two criteria is more important to you?

Left Criterion

Right Criterion

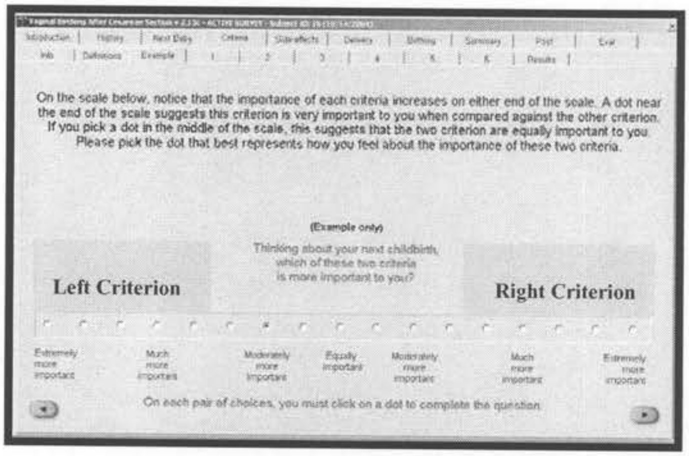
○ ○ ○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○

Extremely more important | Much more important | Moderately more important | Equally important | Moderately more important | Much more important | Extremely more important

On each pair of choices, you must click on a dot to complete the question

Can you also please tell me what this selection means?

4. Show two pictures with "Left Criterion" as choice on left end and "Right Criterion" on the right. One picture has the setting on the left-side "Moderately more important", and the other has the setting on the left-side "Much more important."


Introduction | History | Next Baby | Criteria | Side effects | Delivery | Birthing | Summary | Post | Exit |
Info | Definitions | Example | 1 | 2 | 3 | 4 | 5 | 6 | Results |

On the scale below, notice that the importance of each criteria increases on either end of the scale. A dot near the end of the scale suggests this criterion is very important to you when compared against the other criterion. If you pick a dot in the middle of the scale, this suggests that the two criterion are equally important to you. Please pick the dot that best represents how you feel about the importance of these two criteria.

(Example only)

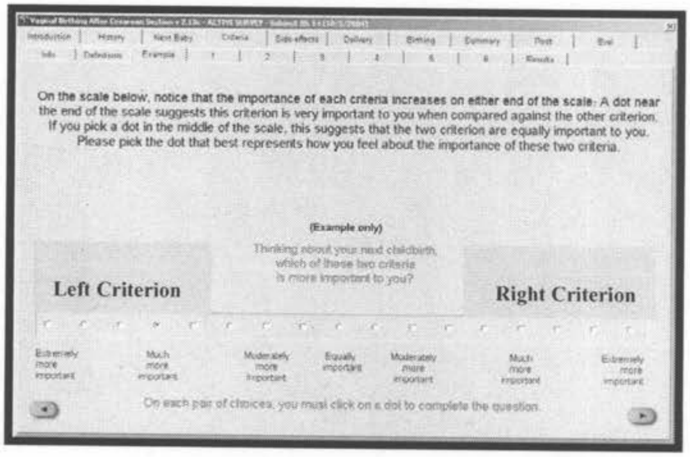
Thinking about your next childbirth, which of these two criteria is more important to you?

Left Criterion **Right Criterion**

○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

Extremely more important | Much more important | Moderately more important | Equally important | Moderately more important | Much more important | Extremely more important

On each pair of choices, you must click on a dot to complete the question.


Introduction | History | Next Baby | Criteria | Side effects | Delivery | Birthing | Summary | Post | Exit |
Info | Definitions | Example | 1 | 2 | 3 | 4 | 5 | 6 | Results |

On the scale below, notice that the importance of each criteria increases on either end of the scale. A dot near the end of the scale suggests this criterion is very important to you when compared against the other criterion. If you pick a dot in the middle of the scale, this suggests that the two criterion are equally important to you. Please pick the dot that best represents how you feel about the importance of these two criteria.

(Example only)

Thinking about your next childbirth, which of these two criteria is more important to you?

Left Criterion **Right Criterion**

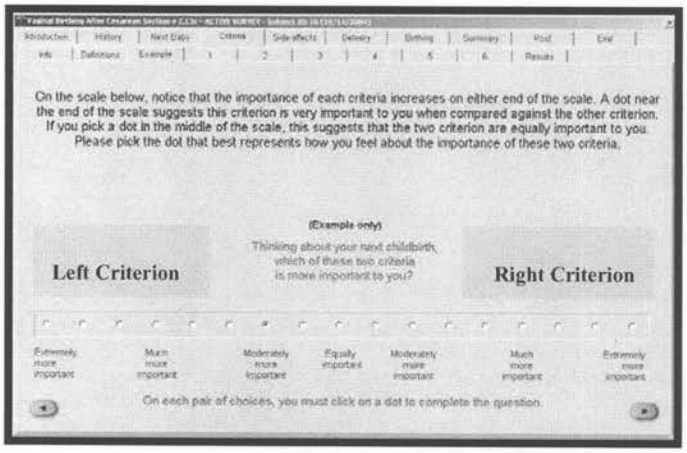
○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

Extremely more important | Much more important | Moderately more important | Equally important | Moderately more important | Much more important | Extremely more important

On each pair of choices, you must click on a dot to complete the question.

What would you say is the difference between these settings?

5. Show two pictures with "Left Criterion" as choice on left end and "Right Criterion" on the right. One picture has the setting on the left-side "Moderately more important", and the other has the setting on the left-side one position to the left of "Moderately more important"



On the scale below, notice that the importance of each criteria increases on either end of the scale. A dot near the end of the scale suggests this criterion is very important to you when compared against the other criterion. If you pick a dot in the middle of the scale, this suggests that the two criterion are equally important to you. Please pick the dot that best represents how you feel about the importance of these two criteria.

(Example only)

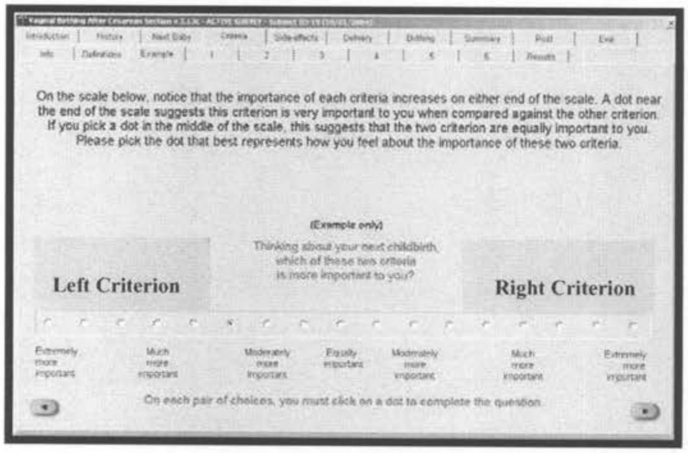
Thinking about your next childbirth, which of these two criteria is more important to you?

Left Criterion — Right Criterion

○ ○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

Extremely more important — Much more important — Moderately more important — Equally important — Moderately more important — Much more important — Extremely more important

On each pair of choices, you must click on a dot to complete the question.



On the scale below, notice that the importance of each criteria increases on either end of the scale. A dot near the end of the scale suggests this criterion is very important to you when compared against the other criterion. If you pick a dot in the middle of the scale, this suggests that the two criterion are equally important to you. Please pick the dot that best represents how you feel about the importance of these two criteria.

(Example only)

Thinking about your next childbirth, which of these two criteria is more important to you?

Left Criterion — Right Criterion

○ ○ ○ ○ ○ ● ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○

Extremely more important — Much more important — Moderately more important — Equally important — Moderately more important — Much more important — Extremely more important

On each pair of choices, you must click on a dot to complete the question.

What would you say is the difference between these settings?

When the subject has finished viewing the pictures of the application, have her answer the following questions concerning her opinion of the Checkoff Scale:

1. *How hard was it to decide where to place your answers?*

2. *Was there the right amount of values on the scale? If not, how many would be just right?*

3. *How easy was it to see your answers on this scale?*

4. *Was the look of this scale appealing? If yes, what makes it appealing? If not, why not?*

5. *Was there the right amount of labels under the scale? If not, would you add more or subtract from those that were there?*

6. *Would you put different labels under the dots? If so; what would they say? Would the scale make sense to you without labels?*

If the subject is using the **checkoff scale first** ask the following question:

7. *The scale uses dots to mark your answers. Is there another way to mark your answers that you think would be better?*

If the subject used the **sliding bar scale first** ask the following two questions:

7. *Now that you've used both dots and a sliding bar, which do you prefer?*

8. *Is there another way to mark your answers that you think would be better than either the dots or the bar?*

For **all subjects** ask the following question:

9. *Would you consider the scale you have just used to be?*

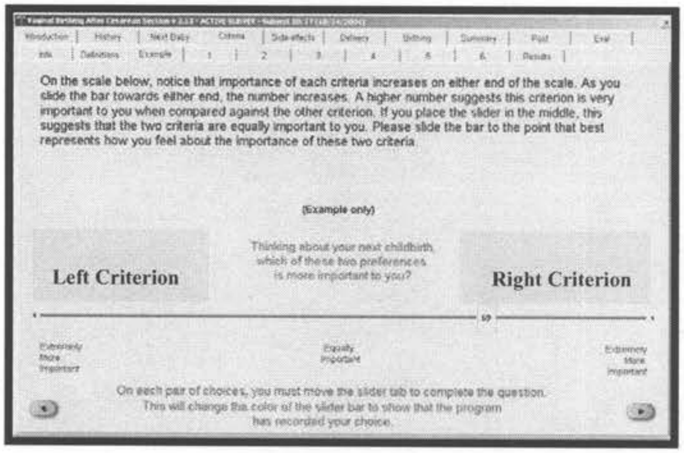
Very difficult
Somewhat difficult
Neither difficult nor easy
Somewhat easy
Very easy

SLIDING SCALE

The following questions with pictures are to determine the subject's understanding the Sliding Scale:

Instruct her to explore its use until she's satisfied that she understands how it works. At that point she may continue on her own. She should answer all application decision questions in the Criteria and Side Effects sections by herself. When she reaches the first Info page of the Delivery section, have her stop. At this point use the questions below.

1. Show picture with "Left Criterion" as choice on left end and "Right Criterion" on the right, slider set at 69.



On the scale below, notice that importance of each criteria increases on either end of the scale. As you slide the bar towards either end, the number increases. A higher number suggests this criterion is very important to you when compared against the other criterion. If you place the slider in the middle, this suggests that the two criteria are equally important to you. Please slide the bar to the point that best represents how you feel about the importance of these two criteria.

(Example only)

Left Criterion

Thinking about your next childbirth, which of these two preferences is more important to you?

Right Criterion

69

Extremely More Important

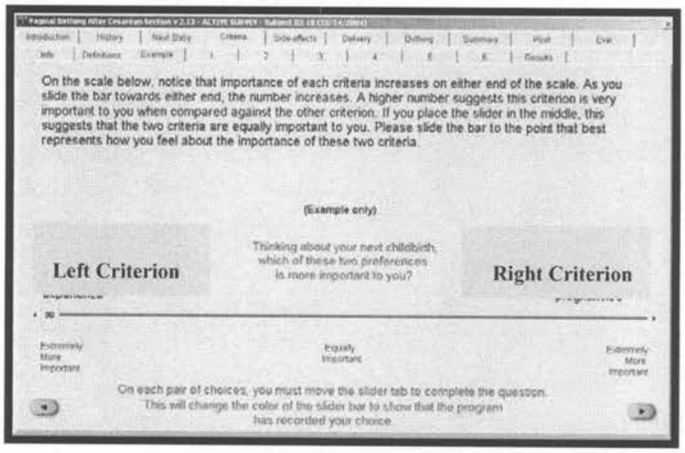
Equally Important

Extremely More Important

On each pair of choices, you must move the slider tab to complete the question. This will change the color of the slider bar to show that the program has recorded your choice.

Suppose you saw another person's answers, and she placed her selection as on this screen. Can you please tell me what this selection means?

2. Show picture with "Left Criterion" as choice on left end and "Right Criterion" on the right, setting at 90 on the left.



On the scale below, notice that importance of each criteria increases on either end of the scale. As you slide the bar towards either end, the number increases. A higher number suggests this criterion is very important to you when compared against the other criterion. If you place the slider in the middle, this suggests that the two criteria are equally important to you. Please slide the bar to the point that best represents how you feel about the importance of these two criteria.

(Example only)

Left Criterion

Thinking about your next childbirth, which of these two preferences is more important to you?

Right Criterion

90

Extremely More Important

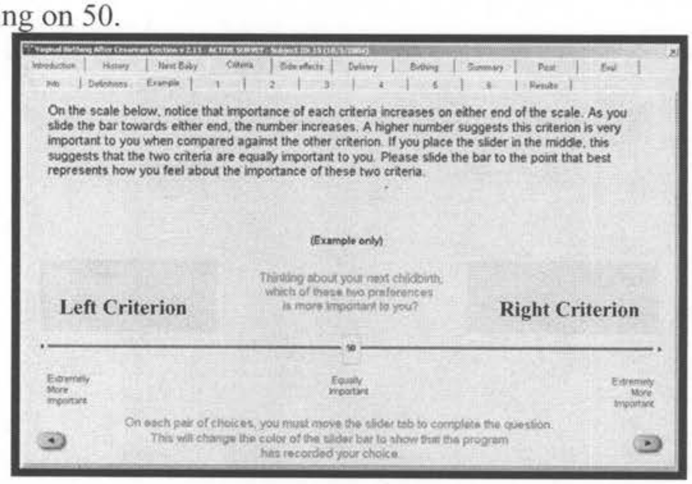
Equally Important

Extremely More Important

On each pair of choices, you must move the slider tab to complete the question. This will change the color of the slider bar to show that the program has recorded your choice.

Can you please tell me what this selection means?

3. Show picture with "Left Criterion" as choice on left end and "Right Criterion" on the right, setting on 50.



On the scale below, notice that importance of each criteria increases on either end of the scale. As you slide the bar towards either end, the number increases. A higher number suggests this criterion is very important to you when compared against the other criterion. If you place the slider in the middle, this suggests that the two criteria are equally important to you. Please slide the bar to the point that best represents how you feel about the importance of these two criteria.

(Example only)

Thinking about your next childbirth, which of these two preferences is more important to you?

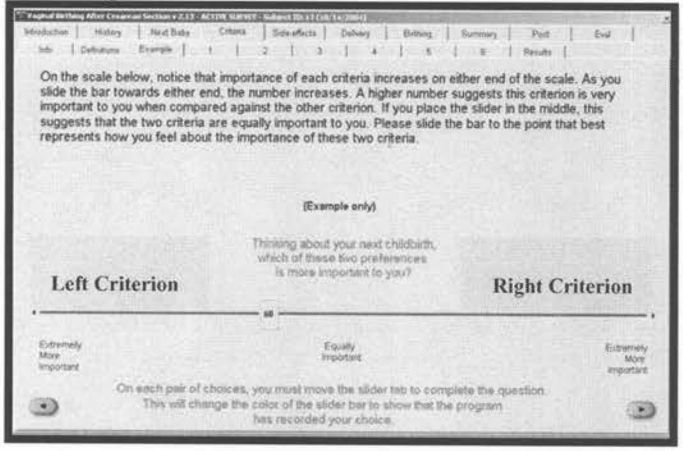
Left Criterion

Right Criterion

On each pair of choices, you must move the slider tab to complete the question. This will change the color of the slider bar to show that the program has recorded your choice.

Can you also please tell me what this selection means?

4. Show two pictures with "Left Criterion" as choice on left end and "Right Criterion" on the right. One picture has the setting on the left-side at 60, and the other has the setting on the left-side at 75.



On the scale below, notice that importance of each criteria increases on either end of the scale. As you slide the bar towards either end, the number increases. A higher number suggests this criterion is very important to you when compared against the other criterion. If you place the slider in the middle, this suggests that the two criteria are equally important to you. Please slide the bar to the point that best represents how you feel about the importance of these two criteria.

(Example only)

Thinking about your next childbirth, which of these two preferences is more important to you?

Left Criterion

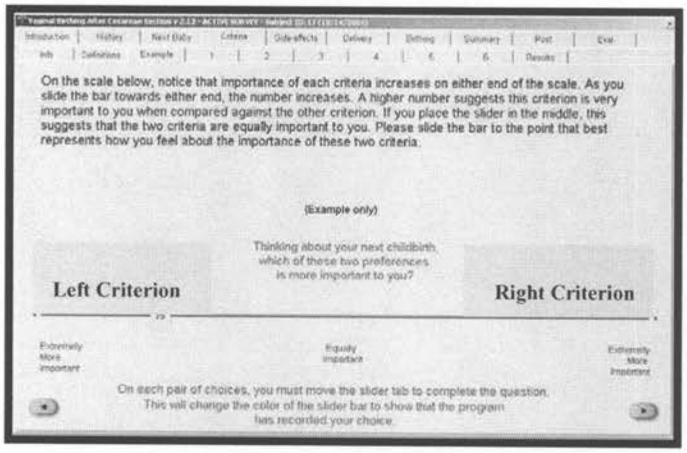
Right Criterion

Extremely More Important

Equally Important

Extremely More Important

On each pair of choices, you must move the slider tab to complete the question. This will change the color of the slider bar to show that the program has recorded your choice.



On the scale below, notice that importance of each criteria increases on either end of the scale. As you slide the bar towards either end, the number increases. A higher number suggests this criterion is very important to you when compared against the other criterion. If you place the slider in the middle, this suggests that the two criteria are equally important to you. Please slide the bar to the point that best represents how you feel about the importance of these two criteria.

(Example only)

Thinking about your next childbirth, which of these two preferences is more important to you?

Left Criterion

Right Criterion

Extremely More Important

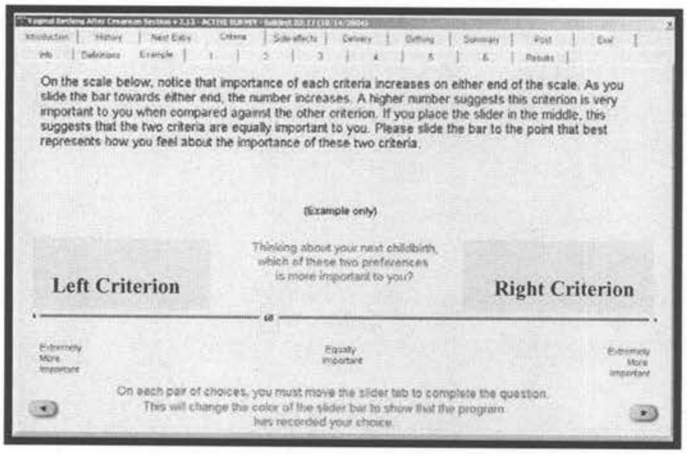
Equally Important

Extremely More Important

On each pair of choices, you must move the slider tab to complete the question. This will change the color of the slider bar to show that the program has recorded your choice.

What would you say is the difference between these settings?

5. Show two pictures with "Left Criterion" as choice on left end and "Right Criterion" on the right. One picture has the setting on the left-side at 60, and the other has the setting on the left-side at 62.



On the scale below, notice that importance of each criteria increases on either end of the scale. As you slide the bar towards either end, the number increases. A higher number suggests this criterion is very important to you when compared against the other criterion. If you place the slider in the middle, this suggests that the two criteria are equally important to you. Please slide the bar to the point that best represents how you feel about the importance of these two criteria.

(Example only)

Thinking about your next childbirth, which of these two preferences is more important to you?

Left Criterion

Right Criterion

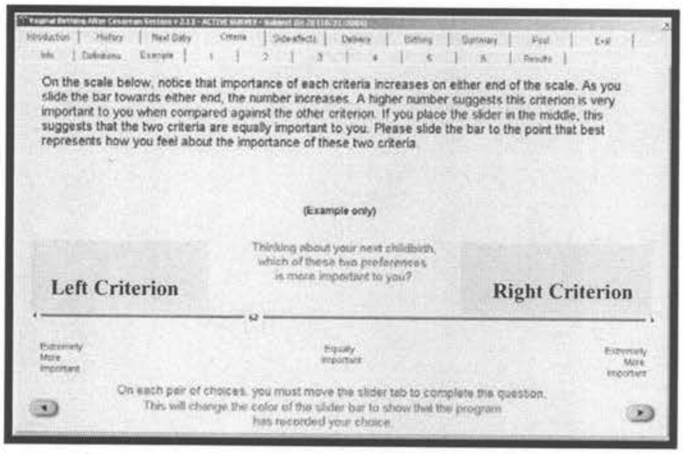
60

Extremely More Important

Equally Important

Extremely More Important

On each pair of choices, you must move the slider tab to complete the question. This will change the color of the slider bar to show that the program has recorded your choice.



On the scale below, notice that importance of each criteria increases on either end of the scale. As you slide the bar towards either end, the number increases. A higher number suggests this criterion is very important to you when compared against the other criterion. If you place the slider in the middle, this suggests that the two criteria are equally important to you. Please slide the bar to the point that best represents how you feel about the importance of these two criteria.

(Example only)

Thinking about your next childbirth, which of these two preferences is more important to you?

Left Criterion

Right Criterion

62

Extremely More Important

Equally Important

Extremely More Important

On each pair of choices, you must move the slider tab to complete the question. This will change the color of the slider bar to show that the program has recorded your choice.

What would you say is the difference between these settings?

When the subject has finished viewing the pictures of the application, have her answer the following questions concerning her opinion of the Sliding Scale:

1. *How hard was it to decide where to place your answers?*

2. *Was there the right amount of values on the scale? If not, how many would be just right?*

3. *How easy was it to see your answers on this scale?*

4. *Was the look of this scale appealing? If yes, what makes it appealing? If not, why not?*

5. *Was there the right range of numbers on the scale? If not, what ranges would you use?*

6. *Would you put different labels than numbers on the bar? If so, what would they say? Would the scale make sense to you without numbers?*

If the subject is using the **sliding bar scale first** ask the following question:

7. The scale uses a sliding bar to mark your answers. Is there another way to mark your answers that you think would.be better?

If the subject used the **checkoff scale first** ask the following two questions:

7. *Now that you've used both a sliding bar and dots, which do you prefer?*

8. *Is there another way to mark your answers that you think would be better than either the dots or the bar?*

For all subjects ask the following question:

9. *Would you consider the scale you have just used to be?*

Very difficult
Somewhat difficult
Neither difficult nor easy
Somewhat easy
Very easy

References

[1] Eds. ISO 9241-11 Ergonomic requirements for office work with visual display terminals (VDTs): Guidance on usability specifications and measures. International Standards Organization. 1998: 9241(11).

[2] Eden KB, Guise JM, Perrin N, Dolan J, Seshadri V, Anderson N, Rosenberg D, McClelland M, Jimison J. Measuring Preferences for Childbirth after Cesarean. In progress.

[3] Eden KB. Proposal for Measuring Preferences for Childbirth after Cesarean. In AHRQ Small Research Grant R03 (PAR-01-040). 2004.

[4] Eden KB, Hashima JN, Osterweil P, Nygren P, Guise JM. Childbirth Preferences after Cesarean Birth: A Review of the Evidence. Birth. 2004 Mar: 31(1), 49-60.

[5] Eden KB, Anderson N, Bhupatiraju RT, Guise JM. Innovative Approach to Measuring Patient Preferences. Proceedings of the Medinfo 2004 Conference, San Francisco. 2004 Sep: 1579.

[6] Nielsen J. Usability Engineering. San Diego: Academic Press, 1993.

[7] Scholtz J. Usability Evaluation. From the National Institute of Standards and Testing: Information Technology Laboratory website. 2004. http://www.nist.gov/itl/iad/IADpapers/2004/Usability%20Evaluation_rev1.pdf

[8] Rubin J. Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests. New York: John Wiley & Sons, Inc., 1994.

[9] Nielsen, J. Guerilla HCI: Using discount usability engineering to penetrate the intimidation barrier. In Bias RG, & Mayhew DJ (Eds.), Cost-Justifying Usability. Boston: Academic Press, 1994. *Also* http://www.useit.com/papers/guerrilla_hci.html.

[10] Nielsen, J. Usability engineering at a discount. In Salvendy G and Smith MJ. (Eds.), Designing and Using Human-Computer Interfaces and Knowledge Based Systems. Amsterdam: Elsevier Science Publishers, 1989. 394-401.

[11] Nielsen, J. Scenarios in Discount Usability Engineering. In Carroll JM. (Ed.), Scenario-Based Design: Envisioning Work and Technology in System Development. New York: John Wiley & Sons, Inc., 1995. 59-83.

[12] Nielsen J, Landauer TK. A mathematical model of the finding of usability problems. Proc. ACM INTERCHI '93 Conference, Amsterdam. 1993 Apr 24-29: 206-213.

[13] Kuutti K. Work Processes: Scenarios as a Preliminary Vocabulary. In Carroll JM. (Ed.), Scenario-Based Design: Envisioning Work and Technology in System Development. New York: John Wiley & Sons, Inc., 1995. 19-36

[14] Saaty, TL. A Scaling Method for Priorities in Hierarchical Structures. Journal of Mathematical Psychology. 1977: 15: 57-68.

[15] Saaty, TL. Analytic Hierarchy Process. New York: McGraw-Hill, 1980.

[16] Saaty, TL. Fundamentals of Decision Making and Priority Theory with the AHP. Pittsburgh, PA: RWS Publications, 1994.

[17] Kocaoglu DF. A Participative Approach to Program Evaluation. IEEE Transactions on Engineering Management. 1983 Aug: EM-30(3): 112-118.

[18] Belton V, Gear T. On a short-coming of Saaty's method of analytic hierarchies. Omega. 1983: 11: 228-230.

[19] Lootsma FA. Numerical scaling of human judgment in pairwise comparison methods for fuzzy multi-criteria decision analysis. In NATO ASI Series F, Computer and System Sciences, Vol. 48, Mathematical Models for Decision Support, Berlin: Springer, 1988, pp. 57-88.

[20] Triantaphyllou E, Lootsma FA, Pardalos PM, Mann SH. On the Evaluation and Application of Different Scales For Quantifying Pairwise Comparisons in Fuzzy Sets. Journal of Multi-Criteria Decision Analysis. 1994: 3: 133-155.

[21] Dolan JG. Involving patients in decision regarding preventative health interventions using the analytic hierarchy process. Health Expectations. 2000: 3: 37-45.

[22] Dolan JG, Frisina S. Randomized Controlled Trial of a Patient Decision Aid for Colorectal Cancer Screening. Medical Decision Making. 2002 Mar-Apr: 22: 125-139.

[23] Shepperd M, Cartwright M. Predicting with Sparse Data. IEEE Transactions on Software Engineering. 2001 Nov: 27(11): 987-998.

[24] Asahi T, Turo D, Schneiderman B. Visual decision making: Using treemaps for the Analytic Hierarchy Process. Presented in Conference companion on Human factors in computing systems a.k.a. CHI '95 Mosaic of Creativity. 1995 May 7-11: 405-406.

[25] Canfora G, Troiano L. The Importance of Dealing with Uncertainty in the Evaluation of Software Engineering Methods and Tools. Proceedings of the 14th international conference on Software engineering and knowledge engineering. 2002 July: 691-698.

[26] Sikorski M. Development consortium: Beyond product usability: User satisfaction and quality management. CHI '00 extended abstracts on Human factors in computing systems. 2000 Apr: 61-62.

[27] Baek S, Liebowitz J, Lewis M. An Exploratory Study: Supporting Collaborative Mulitmedia Systems Design. Multimedia Tools and Applications. 2000 Nov: 12(2-3): 189-208.

[28] Levin MS. Towards Combinatorial Analysis, Adaptation, and Planning of Human-Computer Systems. Applied Intelligence. 2002 Feb: 16: 235-247.

[29] Carroll C, Marsden P, Soden P, Naylor E, New J, Dornan T. Involving users in the design and usability evaluation of a clinical decision support system. Computer Methods and Programs in Biomedicine. 2002 Aug: 69(2): 123-135.

[30] Patel VL, Kaufman DR. Medical Informatics and the Science of Cognition. Journal of the American Medical Informatics Association. 1998 Nov/Dec: 5(6): 493-502.