Evidence for the rapid construction of preference during utility assessments

Jonathan Bennett, M.D.¹
Robert F. Nease, Jr. Ph.D.², Walton Sumner, II, M.D.³
¹Columbia University, College of Physicians and Surgeons New York, NY
²Express Scripts, St. Louis, MO
³Washington University School of Medicine, St. Louis, MO

ABSTRACT

Subjects often construct preferences during the elicitation process. This could have implications for the processes automated utility assessments use to find indifference points. In particular, if subjects frequently shift preferences during elicitation processes, then reversible and irreversible procedures might reach different results. We analyzed series of choices made by two groups of subjects during computerized standard gambles for monocular and binocular blindness, comparing the utility assessed in a reversible search process with the utility that would have resulted from an irreversible search process. A minority of subjects reversed their choices during the assessments. The mean differences between these utilities and the predicted results of an irreversible search were quite small. Consequently, automated standard gamble surveys can terminate quickly, using irreversible searches, with little likelihood of skewing population results. Clinical decision support systems can use simple mechanisms to accommodate infrequent preference reversals, such as restarting a search.

INTRODUCTION

Utilities are quantitative descriptions of the relative values subjects place on health states. These values represent a large fraction of the data required to complete decision and cost-effectiveness analyses. Utility assessments obtain these values from subjects.

Psychologists established that economic and other preferences are constructed during the process of elicitation, based on divergent results from very slightly different assessment techniques. [1] That is, researchers can formulate a task in which most subjects indicate that option A is preferable to option B. In a task with identical information content, a majority of the same subjects indicate that option B is preferable to option A. This is a preference reversal.

Thus, preferences are not simply subjects’ personal values waiting to be discovered, but are constructed on demand. Evidence of preference reversals for health outcomes suggests that a similar, perhaps identical, preference construction process occurs during computer-based utility assessments. [2, 3]

Utility assessments combine a trading metaphor with a search procedure, and must be administered by some interviewer. The metaphors used in the assessment are nearly always a risky standard gamble, a riskless time trade-off, or a currency-based willingness-to-pay. Every metaphor requires a subject to identify a trade-off at the limit of acceptable trades. For instance, a standard gamble often requires that a subject declare indifference between accepting some mortal risk (with a complementary chance of cure) and living in some imperfect state of health, as in the following example.

Choice A: Take a treatment with (1-p) chance of restoring perfect vision, and p risk of causing immediate painless death.

Choice B: Live with blindness in both eyes.

Choice C: Choice A and Choice B are equally attractive.

The subject’s task is to manipulate p until it becomes very difficult to choose between Choice A and Choice B. At that time, the subject may select Choice C, or the interviewer may infer or suggest Choice C. Ideally, subjects could consistently identify a unique value representing the absolute maximum risk they would accept in pursuit of a cure, rather than live with the imperfect state of health, regardless of other assessment details.

The search procedure describes how to manipulate the value of p to identify the indifference point, and often specifies the series of decisions a subject will
make. In a standard gamble, at least four search procedures are sensible. (1) Direct entry requires subjects to enter the acceptable risk, for instance by typing, clicking on a visual analog scale representing risk, dialing a value with a control, or selecting a value from a list. (2) Titration would give subjects a series of opportunities to declare indifference as risk steadily rises (or falls). The point where the subject first rejects an incremental increase (or decrease) is close to the indifference point. (3) Ping-pong combines and alternates titration rising from zero risk and falling from 100% risk. (4) Bisecion usually offers zero risk, 100% risk, then bisects ranges of risk until the subject becomes indifferent.

Bisecion offers several advantages as a search process. First, the initial offers validate responses anywhere in a broad range. Ping-pong shares this advantage, while titration could bias results toward the starting point. Second, bisecion and ping-pong encourage subjects to explicitly consider a wide range of responses. In contrast, direct entry does not usually encourage explicit consideration of slightly more risky scenarios. Third, bisecion is the most efficient search process if no information is available on the distribution of utilities in a population of subjects. Even if the distribution of utilities is known, a customized search process risks biasing the results to save only a few choices per assessment.

The search procedure itself may affect subjects’ preference construction process, and therefore the results obtained. [1-5] In addition to leaving us uncertain about which search procedure is best, this fact raises questions about terminating preference assessments. If preference construction occurs very early in utility assessments, then the interviewer can terminate assessments with the first indication that p is close to the indifference point. Conversely, if construction occurs late in utility assessments, then the interviewer should allow large changes at any time in the process. Anecdotal reports suggest that some subjects have insights very late in an assessment process, and may even decide not to pursue the medical intervention that brought them to the attention of an investigator.

Concerns about human interviewer bias and the desire to conduct reproducible assessments led several groups to develop automated preference assessment tools, in effect creating software interviewers. [6, 7] Paper-based surveys are a very tempting approach to conducting affordable preference surveys. [8] Human interviewers can solve new assessment problems in real time, and intelligently respond to subjects’ insights. There is no similar problem-solving expectation of paper-based interviews. Computer-based preference assessment tools fall between these extremes. A software interviewer might be expected to gracefully react to shifts in preferences and recognize patterns of behavior that suggest indifference.

A software interviewer may allow subjects to express a newly constructed preference whenever it becomes evident to the subject. We can describe these interviews as being reversible. The Impact preference assessment program [7] uses reversible processes, including ping-pong and titration methods with undo commands, as well as direct entry methods. The U-Titer preference assessment program includes a reversible bisecting search process, in which the interviewer approximately bisects ranges of risk until the subject is either explicitly or implicitly indifferent to a risk. [6] Indifference is inferred when a subject confirms a recent decision to accept or reject a risk. When a subject changes responses to a risk during the assessment, the U-Titer shifts the bisecion range to locate the new indifference value. U-Titer maintains stacks of previous responses, and pops values off the stacks to find new bisecion ranges.

We can describe some software interviewers as irreversible. Irreversible interviewers terminate an assessment when a subject switches the direction of the search, even though the switch could indicate a desire to return to a previously rejected preference value. An irreversible interviewer could therefore misinterpret responses if the subject constructs preferences during the assessment. For instance, a declining risk titration search may offer sequentially lower risks in standard size steps, and terminate with the first risk accepted by the subject. This search pattern might first offer a 100% risk, which the subject normally rejects. It could then offer 95%, then 90%, until the subject accepts some level of risk. The indifference point used to calculate utilities would be halfway between the last risk rejected and the first risk accepted. However, imagine that during this process the subject reaches 50% risk, then constructs a new preference such that the acceptable risk exceeds 90%. Accepting the 50% risk causes an irreversible software interviewer to report indifference at 52.5% risk. Irreversible search strategies are easier to program, but might bias results or introduce large random errors. Interviewers might also benefit from an estimate of the frequency of major preference revision late in assessment processes.

**METHODS**

We analyzed records of subjects who participated in two utility assessment interviews. Group I consisted
of 299 ophthalmology patients who participated in a survey regarding visual function. Group II consisted of 114 parents of children seen in an urban emergency ward who participated in a survey regarding outcomes of occult bacteremia.

Both groups answered standard gamble questions regarding binocular and monocular blindness early in their respective surveys. Both surveys were constructed and conducted using U-Titer II. Standard gambles offered a cure with 0% risk of death, then 100% risk, then used a successive bisection routine to determine indifference points. When the bisection reached a resolution of 1%, the subject either confirmed or rejected a previous boundary value. The reversible software interviewer repeats previous risk offers if the user rejects the old boundary. We used U-Titer II records of each subject’s choices in each assessment to reconstruct the sequence of risk offers. These records also indicate errors that prematurely terminate interviews, such as repeatedly rejecting a risk-less procedure that cures blindness.

A sequence of ten or fewer offers contains no reversals of earlier choices. After the ninth offer, the bisection range diminishes to a 1% interval. If the tenth choice confirms a boundary of the 1% interval, the program infers an indifference point within the 1% range. If the tenth choice does not confirm the boundary, the indifference point has changed and lies in a different range. Because the subject is confined to binary choices and can not undo a choice, finding this new indifference point requires more than ten choices and at least one reversal. Thus, a subject who constructs a new utility is likely to reverse an earlier choice and make more than ten choices.

We defined the irreversible utility value as the value at the conclusion of the ninth offer (or the last offer, if a subject made fewer than nine choices). This is the utility that an irreversible bisection routine would have reported. We defined the reversible utility value as the ultimate result of the assessment. When no reversals occur, these utilities are equal.

Finally, we calculated the time required to support reversible searches as the difference between the time at the beginning of the tenth offer and the time at the conclusion of the last offer in a search pattern. When no reversals occur, this time difference is zero. We looked at effects reversible searches had on the mean utility, median utility, individual utilities, and assessment times.

RESULTS

Table 1 summarizes demographic data on the two groups. Group I subjects were more likely to be older, white, college graduates, and to have higher income than subjects in group II (p<0.01, t-test or chi-square as appropriate). Subjects in both groups were predominantly female, but more males were present in group I.

<table>
<thead>
<tr>
<th>Group</th>
<th>I</th>
<th>II</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>299</td>
<td>114</td>
</tr>
<tr>
<td>Median Age (25th - 75th percentiles)</td>
<td>64 (48-72)</td>
<td>26 (22-33.4)</td>
</tr>
<tr>
<td>Female (%)</td>
<td>59</td>
<td>92</td>
</tr>
<tr>
<td>African-American</td>
<td>6</td>
<td>64</td>
</tr>
<tr>
<td>College Graduates</td>
<td>37</td>
<td>23</td>
</tr>
<tr>
<td>Income &gt;$50,000</td>
<td>45</td>
<td>23</td>
</tr>
</tbody>
</table>

In the binocular blindness standard gamble, 12% of subjects in both groups reversed a choice. In the monocular blindness standard gamble, 18% of subjects in group I and 9% of subjects in group II reversed a choice. (Table 2). For those individuals who reversed a choice, the difference between irreversible and reversible utilities are displayed in histograms (Figs. 1-4). In each utility assessment about half of reversals resulted in utility differences of less than or equal to +/- 0.05. Table 2 shows the median and means of the differences in utilities, as well as the median and mean of the extra time required for the assessments with reversible searches.

<table>
<thead>
<tr>
<th>Table 2.</th>
<th>Group I</th>
<th>Group II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blindness affects:</td>
<td>both eyes</td>
<td>one eye</td>
</tr>
<tr>
<td>Reversed search pattern’s(%)</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>Mean of utility difference</td>
<td>.01</td>
<td>.02</td>
</tr>
<tr>
<td>Median of utility difference</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Range of utility differences</td>
<td>.01-.5</td>
<td>.01-.74</td>
</tr>
<tr>
<td>Mean of time differences (sec)</td>
<td>9.72</td>
<td>17.33</td>
</tr>
<tr>
<td>Median of time difference (sec)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mean time extension of reversed searches (sec)</td>
<td>83.7</td>
<td>101.1</td>
</tr>
<tr>
<td>Median time extension of reversed searches (seconds)</td>
<td>66.5</td>
<td>81</td>
</tr>
</tbody>
</table>
The statistical differences between irreversible and reversible utilities are quite small. The mean of the utility differences for the different groups ranges from 0.01 to 0.07. The median of the utility differences ranges from 0 to .04. However, for individual subjects the difference between reversible and irreversible utilities was as large as 0.75.

Reversible searches lengthened the mean assessment time by 9.72 to 17.47 seconds, but did not change the median interview time. For the subset of individuals who reversed choices, reversible searches lengthened the mean interview times by 83.7 to 181.09 seconds, and median times by 66.5 to 113 seconds.

DISCUSSION

The finding that individuals construct preference values during elicitation could have important implications for terminating computer-based search processes used to measure preferences. Fortunately, our results suggest that this is not a major problem for survey data.

Subjects who made lengthy searches for an indifference point could have been constructing preferences during their search. Choice reversals consistently occurred in these four assessments, but only among a minority of individuals. With subjects from two distinct demographic groups, assessing two outcomes of quite different severity, only 10-20% reversed choices.

Furthermore, software support for preference reversals had very little impact on the sample mean or median utility. One reason for this lack of difference was that even when choice reversals occurred, the final, reversible utility was usually within +/- 0.05 of the irreversible utility. A second reason was the relatively symmetric distribution of differences. This implies that computer-based surveys seeking to describe statistical distributions of utilities do not require reversible search procedures.

These findings suggest that the construction of preference generally occurs very quickly. If preferences were constructed late in these bisection searches, we
would have expected more choice reversals, and more disparity between irreversible and final utility values. The apparent speed of preference construction gives us some reason to hope that this stability will be found with search procedures other than bisection.

Although overall results were stable, a minority of subjects made a large shift in utility during the assessment process. Irreversible searches would have resulted in errors as large as $+/−0.75$ in these assessments. Any preference assessment intended to inform a decision facing an individual patient should obviously be reversible. Person-to-person interviews are reversible in the sense that the subject can verbally correct mistakes. Utility assessments in live Internet decision support could usually be restarted even if no other error correction mechanism were offered. Therefore flexibility is unlikely to be an issue for live preference assessments. Nevertheless, data collected with an irreversible technique could confuse an analyst who wanted to solve a decision model for each subject in a utility survey. If some of the utilities were in error to this degree, a few inferred decisions would be quite wrong. This could result in expensive errors. For instance, a policy maker might conclude that a minority of subjects strongly desires access to a technology that no one really wants.

Limitations
There are limitations to these data. First, they were not collected for this study originally. We have tested only the bisection search procedure directly, and can not describe other procedures. Choice reversals could result when initially confused subjects finally understand the assessment task. Therefore, these results are not direct evidence of preference construction during assessments. Finally these data involve only utility assessments about blindness. However, in data from smaller populations assessing heart failure and migraine headaches we have seen a similar frequency of reversed choices (about 12%).

CONCLUSIONS
It appears from our data that reversible search strategies can produce utilities that differ markedly from the first utility found by an irreversible search strategy, but only for a small minority of individuals. Reversable search strategies have almost no effect on the central tendency of moderately sized samples. This finding may be helpful in settings where reversible utility assessments are impractical.

References