Preferences for Health Outcomes
Comparison of Assessment Methods

J. Leighton Read, M.D.,
Robert J. Quinn, Ph.D.,
Donald M. Berwick, M.D.,
Harvey V. Fineberg, M.D., Ph.D.,
and Milton C. Weinstein, Ph.D.

This study compared standard gamble (SG), time trade-off (TTO), and category scaling (CS) methods for assessing preferences among hypothetical outcomes of coronary artery bypass surgery. High correlations among assessment methods, as found in some previous studies, do not assure the absence of systematic differences in ratings obtained by different methods. This study used analysis of variance to test for differences among the three assessment methods.

Questionnaire responses were obtained from 67 of 109 physicians participating in a postgraduate course on clinical decision making, following a lecture and workshop on utility theory. SG and CS were used to rate multivariate combinations of angina (none, moderate, and severe) and survival (0, 5, and 10 years); and SG, TTO, and CS were used to rate univariate outcomes with angina (none, moderate, and severe) for the remainder of their life expectancy.

SG ratings were higher than TTO ratings, which were higher than CS ratings ($p < 0.001$ for all comparisons). Multivariate responses revealed a significant interaction between angina and survival dimensions using CS, but not using SG. We conclude that these methods are not interchangeable and that differences between SG and CS require a more complex explanation than differences in attitude toward risk. (Med Decis Making 4:315-329, 1984)

Introduction

Quantitative models of medical decisions require explicit estimates of probabilities and focus attention on the relative values of health outcomes such as survival time, symptoms, and physical limitations. The medical and epi-
Figure 1. Preference assessment techniques: Lottery method.
1. Pick end points.
2. Find $p$ for which indifference exists between lottery on best and worst outcomes versus certain intermediate outcomes.

Demiological literatures are replete with information about frequencies that may be pertinent to probability estimates. However, a similar repository of information on how people value the relative importance of different health outcomes is not available.

Economic and psychological theories attempt to explain the way people structure their preferences. Although these disciplines provide frameworks for exploring preferences, a great deal of controversy surrounds the choice among methods for describing individuals' preferences at the level of precision required by quantitative models of decision making.

The decision scientists who adhere most closely to the economic theories often argue that the "standard gamble" method for quantifying values has advantages over other methods because this technique is derived from intuitively appealing axioms of utility theory, as developed by von Neumann and Morgenstern [1] and extended by Savage [2] and by Keeney and Raiffa [3]. The standard gamble method requires that people choose between hypothetical lotteries [4,5]. When the decision problem under study involves...
uncertainty, the standard gamble technique is said to be more appropriate than methods that do not incorporate the element of risk in the assessment task [4,6,7]. Figure 1 illustrates a standard gamble designed to elicit the value, or utility, of one possible outcome of coronary artery disease.

Although decision scientists with backgrounds in experimental psychology have used the standard gamble, they generally advocate the use of other techniques, adapted from research in psychophysics and attitude scaling [8–10]. Psychologists have favored assessment techniques in which the subject must either rank outcomes or represent them on a numerical, verbal, or graphical scale. Relatively simple direct rating scales, as depicted in Figure 2, have been used both to assist decision makers with complex multiattribute problems [11] and to obtain population based preferences for constructing a general health status index [12]. The time trade-off technique shown in Figure 3 is another of the so-called direct methods for preference assessment [10,13]; unlike category scaling, it can be justified by the axioms of utility theory, under certain conditions [14].

The theoretical advantages of a particular assessment method also may depend on whether it is to be used to aid decisions by individuals or by groups. For program evaluation and other forms of group decision making there is no well accepted framework to guide the aggregation of individual preferences.

**Comparability of Assessment Methods**

In choosing a preference assessment technique to aid decision making, it would be helpful to know whether lottery-based methods such as the standard gamble (SG) and more direct methods such as category scaling (CS) or time trade-off (TTO) yield different results. Since risk is an important part of the standard gamble task, preferences assessed with SG may depend, in part, on the subjects' attitudes toward risk. Because risk is not present in the stimuli usually presented in CS or TTO assessment methods, these methods may yield preference ratings substantially different from those obtained with SG.

Responses to preference questions can be strongly influenced by the way the questions are formally presented or “framed.” For example, lotteries with outcomes framed as losses result in risk seeking, while equivalent lotteries with outcomes framed as gains result in risk aversion [15–17]. Survey researchers have shown that small changes in question wording can influence responses in some situations, though not in others [18–20]. Since the structure and wording of questions used in SG, CS, and TTO methods are so different, it would not be surprising to find that they produce different answers.

A few comparative studies have found that these methods do produce different results. Torrance [13] studied SG, CS, and TTO in 43 randomly
selected college alumni who rated six health state scenarios. The (uncorrected) Pearson product–moment correlation between SG and TTO for these 258 ratings was 0.65, which Torrance interpreted as satisfactory, in contrast to the lower SG–CS correlation of 0.36. Wolfson et al. reported correlations between ratings with SG and TTO of 0.84, between SG and CS of 0.76, and between TTO and CS of 0.89 in a group of stroke patients [21]. They noted, however, significant differences between SG and TTO for 17 of the 35 health states rated and between SG and CS for 33 of 35 states rated. TTO and CS differed for only 8 of 35 health states. Krzysztofowicz reported systematic differences between 56 responses and judgments using a value assessment procedure termed the “exchange method” [22]. Also, Quinn found significant differences between CS and SG responses in a medical decision making task, using college students as subjects [23].

Not all investigators have observed such intermethod differences. Fischer [24,25] and von Winterfeldt [26] observed a high degree of convergence (Kendall’s tau correlation coefficients ranging from 0.85 to 0.95) between SG and direct evaluation methods. Patrick, Bush, Kaplan, and their colleagues [27,28], who have used category scaling techniques in the construction of a general health status index, found CS ratings for levels of dysfunction were comparable to those produced by a population trade-off procedure they called the “equivalence method.” They concluded that the CS

![Category Rating Technique](image)

Figure 2. Preference assessment techniques: Category rating technique.
1. Pick a scale (e.g., 0–100).
2. Assign best and worst outcomes to top and bottom of scale.
3. Fill in values of intermediate outcomes.
Figure 3. Preference assessment techniques: Time trade-off technique.
1. Pick a time horizon and reference state (e.g., 10 years and perfect health).
2. Find what fraction $Q$ of years in the best state is equivalent to 10 years in each other state.

...method was comparable to SG, because their equivalence method was theoretically identical to TTO, which Torrance found to be correlated with SG.

In another paper, Kaplan et al. [29] argue that Torrance’s low CS-SG correlation (0.36) can be traced to the use of a narrow range of data points and to the subjects’ poor comprehension of the CS task and stimulus items. The latter argument, however, would not explain the substantial CS-SG differences found by Quinn [23] in a study where high ($r > 0.85$) test-retest reliabilities were observed for each method.

**PURPOSE OF THE STUDY**

Correlation is a poor way to detect systematic differences between methods of preference assessment; it does not allow rejection of hypothesized convergence between two sets of responses [30].

Analysis of variance does allow this hypothesis testing. The present study used analysis of variance to test the null hypothesis that SG, CS, and TTO produce convergent preference scales. Preference judgments were elicited from physicians who had received training in preference assessment methodology.

**Methods**

The subjects in this study were physicians enrolled in a three-day postgraduate course on clinical decision making at the Harvard School of Public Health. After two days of lectures and workshops on decision analysis and the theory of test selection and interpretation the participants had a lecture on utility assessment in which the “basic reference lottery method” was introduced and explained [5]. Then they participated in a small group workshop led by faculty members. The workshop concerned a hypothetical patient facing the decision whether to undergo coronary artery bypass sur-
Figure 4. Decision tree for coronary bypass surgery.

Surgical outcomes and combinations of angina levels and survival times were evaluated as if participants were that patient. This model was based on published decision analyses [31,32], though considerably simplified for the didactic purposes of the workshop.

Participants first ranked the seven possible multivariate outcomes from most to least desirable, and then assigned each outcome a rating on a 0 to 100-point category scale (with the anchors fixed at the subjects' own best and worst outcomes, as shown in Figure 2). Then they constructed a basic
reference lottery to evaluate the second best outcome on their list. Figure 1 shows such a lottery. The subjects’ utility for this outcome (or preference rating, on a 0–1 scale) was the probability of the best outcome (p) at which they would be exactly indifferent between choosing the lottery and obtaining the second best outcome for sure. After group discussion and questions about the technique, each participant evaluated the remainder of the seven hypothetical outcomes with variable probability lotteries, using their own best and worst outcomes as anchors.

Following the workshop, participants were asked to complete an anonymous questionnaire reporting the values they had given to the seven multivariate outcomes using each assessment method (ranking, category scale, and standard gamble). In addition, they were asked to evaluate hypothetical univariate outcomes in which they would suffer from either moderate or severe angina for the remainder of their own life expectancy. SG, CS, and a time trade-off procedure (Figure 3) were used for these univariate outcomes.

**Results**

Sixty-seven of the 109 course registrants (61%) returned the questionnaire. Since we observed that some of the registrants were absent for the final day of the course, when the questionnaire was administered, the figure represents the lower bound of our response rate. Seven questionnaires were incomplete and were omitted from the analysis. Respondents ranged in age from 27 to 56 (mean age 39). Nine of the 60 subjects (15%) were women. Thirty-six percent considered their primary activity to be patient care, 25 percent teaching, and 16 percent research. Twenty-three percent were primarily administrators or indicated two or more primary activities.

**Single-Attribute Preferences**

Physicians used the three assessment methods to evaluate the single-attribute outcomes of experiencing moderate or severe angina with survival time specified as “the rest of your life.” Scale anchors were labeled “immediate death” and “perfect health” for all methods. Correlational analysis of moderate and severe angina ratings, aggregated across all 60 subjects, revealed a moderately strong relationship among response sets produced by all three methods (2 responses × 60 subjects = 120 observations). The Pearson product-moment correlation between SG and CS responses was 0.63 (p < 0.01). Between SG and TTO it was 0.65 (p < 0.01), and between CS and TTO it was 0.65 (p < 0.01). It must be emphasized, however, that since these statistics reflect variability across subjects, as well as methods, they cannot be compared readily with those of studies using different subjects and sample sizes (e.g., [13]).

Analysis of variance revealed a pattern of systematic differences among preference methods. Mean scores for each outcome and method are shown
Table 1. Mean Ratings Across Methods for Single-Attribute Health Outcomes

<table>
<thead>
<tr>
<th>Assessment method</th>
<th>Moderate angina</th>
<th>Severe angina</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>71.8</td>
<td>35.4</td>
</tr>
<tr>
<td>SG</td>
<td>90.3</td>
<td>70.7</td>
</tr>
<tr>
<td>Significance</td>
<td>( p &lt; 0.001 )</td>
<td>( p &lt; 0.001 )</td>
</tr>
<tr>
<td>CS</td>
<td>71.8</td>
<td>35.4</td>
</tr>
<tr>
<td>TTO</td>
<td>83.2</td>
<td>53.3</td>
</tr>
<tr>
<td>Significance</td>
<td>( p &lt; 0.001 )</td>
<td>( p &lt; 0.001 )</td>
</tr>
<tr>
<td>SG</td>
<td>90.3</td>
<td>70.7</td>
</tr>
<tr>
<td>TTO</td>
<td>83.2</td>
<td>53.3</td>
</tr>
<tr>
<td>Significance</td>
<td>( p &lt; 0.005 )</td>
<td>( p &lt; 0.001 )</td>
</tr>
</tbody>
</table>

\( ^a \) SG = standard gamble; TTO = time trade-off; CS = category scaling.  
\( ^b \) All significance levels pertain to \( F \)-statistics from one-way ANOVA.

in Table 1. For both angina levels the SG ratings were higher than the CS and TTO values (one-way ANOVA, \( df = 1/118, p < 0.001 \) for each comparison, \( n = 60 \)). That is, moderate and severe angina were seen as closer to perfect health (relative to death) by SG than by CS or TTO. Mean TTO responses to both angina items fell between the CS and SG means and were significantly different from both. Analysis of data for individual subjects revealed that this ordinal relationship (SG > TTO > CS) held for 43 of the 60 respondents. Only three subjects produced category ratings for angina that were greater than the corresponding SG responses. Thus the differences among the three assessment methods, shown in Table 1, held at both the group and individual levels.

**MULTIATTRIBUTE PREFERENCES**

*Ordinal Analysis.* Not all subjects produced identical preference orders for the seven multiattribute outcomes in Figure 4. Most strikingly, 11 (18%) of the subjects failed to rate death as the worst outcome in at least one of the three response modes. Nine of these subjects responded in this way with all three assessment methods. Seventeen (28%) of the 60 subjects answered in at least one of the response modes that they would prefer to live five years rather than 10 if they had severe angina. Of these, two preferred five years over 10 years in the moderate condition as well. One indicated that five years with moderate angina was preferred to five years with no angina, possibly due to error or confusion. With these exceptions, the physicians preferred less angina to more and longer survival to shorter.

*Interval Analysis.* In order to aggregate interval scales across subjects...
using the same scale endpoints, our sample was divided into two groups: those who used surgical mortality as the worst outcome \((n = 49)\) and those who did not \((n = 11)\). The larger sample was used in the main analysis of CS and SG that follows. TTO responses were not elicited for the multivariate (angina and survival) outcome because the method scales preferences for angina in terms of survival.

A moderately high Pearson product–moment correlation coefficient \((r = 0.56, p < 0.01)\) was obtained between the CS and SG ratings of multiattribute outcomes \((n = 49 \text{ subjects} \times 5 \text{ scale values} = 245 \text{ observations})\). As in the univariate condition, however, the mean SG ratings for these outcomes were significantly greater than the corresponding CS means (Table 2). Moreover, for each health outcome the SG values equaled or exceeded the CS values for more than 90 percent of the 49 respondents. The group effect, therefore, does not appear to be due only to a large difference on the part of a small number of subjects.

Of the 11 subjects who did not rate surgical mortality as the worst outcome, seven used the same scale anchors for both CS and SG, thus allowing the analysis performed on the main group to be carried out separately on this subsample. These analyses revealed that SG ratings were significantly greater than CS values for two of the five intermediate outcomes \((p < 0.05)\) and in the same direction (though not statistically significant) for the other three. Given the small sample size, it is reasonable to conclude that the effect of assessment method holds for both sets of subjects.

**Method Variance and Stimulus Characteristics.** We examined the relative importance of angina, survival, and their interaction by performing

---

Table 2. Mean SG and CS Ratings for Multiattribute Health Outcomes

<table>
<thead>
<tr>
<th>Survival</th>
<th>Angina</th>
<th>Mean CS</th>
<th>Mean SG</th>
<th>Significance level</th>
<th>Number of subjects for whom SG ≥ CS ((n = 49))</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 yrs</td>
<td>None</td>
<td>100</td>
<td>100</td>
<td>(p &lt; 0.001)</td>
<td>47</td>
</tr>
<tr>
<td>10 yrs</td>
<td>Moderate</td>
<td>82</td>
<td>94</td>
<td>(p &lt; 0.001)</td>
<td>44</td>
</tr>
<tr>
<td>5 yrs</td>
<td>None</td>
<td>42</td>
<td>71</td>
<td>(p &lt; 0.001)</td>
<td>47</td>
</tr>
<tr>
<td>5 yrs</td>
<td>Moderate</td>
<td>70</td>
<td>91</td>
<td>(p &lt; 0.001)</td>
<td>47</td>
</tr>
<tr>
<td>5 yrs</td>
<td>Severe</td>
<td>49</td>
<td>84</td>
<td>(p &lt; 0.001)</td>
<td>46</td>
</tr>
<tr>
<td>Surgical mortality</td>
<td></td>
<td>24</td>
<td>62</td>
<td>(p &lt; 0.001)</td>
<td>0</td>
</tr>
</tbody>
</table>

CS = category scaling.
SG = standard gamble.

\(^{a}\) One-way analysis of variance of difference between CS and SG; df = 1/47.
Table 3. ANOVA Results for Two-by-Two Subset of Health Outcomes

<table>
<thead>
<tr>
<th>Source of variance</th>
<th>df</th>
<th>F</th>
<th>$\Omega^{2b}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STANDARD GAMBLE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate versus severe angina</td>
<td>1</td>
<td>64$^c$</td>
<td>22%</td>
</tr>
<tr>
<td>10 versus 5 years survival</td>
<td>1</td>
<td>63$^c$</td>
<td>4%</td>
</tr>
<tr>
<td>Angina $\times$ survival interaction</td>
<td>1</td>
<td>&lt;1.0</td>
<td>0%</td>
</tr>
<tr>
<td>Pooled error terms$^d$</td>
<td>192</td>
<td></td>
<td>74%</td>
</tr>
<tr>
<td>Total</td>
<td>195</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td><strong>CATEGORY SCALING</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate versus severe angina</td>
<td>1</td>
<td>298$^c$</td>
<td>34%</td>
</tr>
<tr>
<td>10 versus 5 years survival</td>
<td>1</td>
<td>109$^c$</td>
<td>22%</td>
</tr>
<tr>
<td>Angina $\times$ survival interaction</td>
<td>1</td>
<td>38$^c$</td>
<td>2%</td>
</tr>
<tr>
<td>Pooled error terms$^d$</td>
<td>192</td>
<td></td>
<td>42%</td>
</tr>
<tr>
<td>Total</td>
<td>195</td>
<td></td>
<td>100%</td>
</tr>
</tbody>
</table>

$^a$ Analysis based on two-factor, completely repeated design [33].
$^b$ $\Omega^{2}$ may be interpreted as percent of total responses variance accounted for by each effect [33].
$^c$ $p < 0.001$, two-tailed test.
$^d$ Includes all effects $\times$ subjects interactions.

analysis of variance on a four-cell subset of health outcomes (five and ten years survival by moderate and severe angina). The results, presented in Table 3, may be summarized as follows:

1. Both angina severity and survival time influenced responses for both assessment methods, as reflected by the large $F$-statistics for these outcomes.

2. The interaction between angina and survival was significant only for the CS condition; that is, only for CS ratings did the value of five additional years of life depend upon the severity of angina. While the $F$-statistic (38) for this interaction is significant, its contribution was much less important for predicting responses than was angina or survival alone. Another way to describe this contribution is with the $\Omega^{2}$ [33] statistic, a measure of each term's percent contribution to explaining subjects' ratings. As shown in the table, the angina $\times$ survival interaction accounted for only two percent of the variability in category scale responses.

3. The variability of responses among subjects (error terms in Table 3) is considerably greater for SG than for CS.

4. The relative importance of angina versus survival depends on assessment method, at least for these four combinations of angina and survival. Note in Table 3 that for SG the difference between severe and moderate angina contributes nearly six times as much variance to
responses as the difference between five and 10 years of life (22% versus 4%), whereas the corresponding ratio for CS is only 1.5. Therefore, the SG–CS effect in the multiattribute problem cannot completely be reduced to the same effect seen in the single-attribute problem: the way in which angina and survival are evaluated relative to one another also seems to be influenced by assessment method.

Discussion

The principal finding of this study is that, for a group of 60 physicians, category scaling, standard gamble, and time trade-off procedures produced different scale values for outcomes of a clinical problem. In univariate and multivariate tasks physicians saw intermediate health outcomes as less aversive, relative to the scale anchors, when responding with SG than with CS. The time trade-off values fell between the SG and CS ratings and were significantly different from both.

Previous studies often relied on high correlations to demonstrate comparability across methods of assessment. As shown in our analysis, high correlations can coexist with systematic differences between sets of scale values.

The difference between TTO and SG contrasts with the findings of Torrance [13]. A number of factors could account for these discrepancies, including differences in subject characteristics, type of medical situation assessed, and details of the assessment procedures. TTO uses the dimension of life years as a metric to scale single-attribute health states such as angina levels, while SG does not. If the value for future life years is discounted relative to that for the current year (a common result), then one would expect SG and TTO to produce differing responses.

The conclusions of this study are bolstered by the strength of the observed effect and by the fact that our medically sophisticated subjects had received some training in utility analysis. However, our results agree with those of previous research on a sample of college students [23], in which CS and SG responses did pass tests of reliability (test-retest $r_{SG} = 0.85$; $r_{CS} = 0.95$) and model validation for each individual.

Several psychological mechanisms might explain the effect of assessment method on preference responses. These explanations are not mutually exclusive:

1. **Response spreading.** Experimental evidence indicates that one factor affecting CS responses is a desire to use all categories of the scale equally often [34]. Thus in our case, subjects might have wanted to space their intermediate outcomes over the entire 100-point scale, even if the "true" values were bunched at one end. This explanation does not account, however, for the SG–TTO differences also found in this study.
2. Influence of attitude toward risk. The SG method explicitly introduces risk, while the CS method does not. Many decision theorists would explain the CS–SG differences in terms of an aversion or attraction to gambling per se, independent of specific outcome attributes [4,6]. Our finding of an interaction between the CS–SG effect and outcome attribute suggests, however, that a more complex explanation is needed. For example, one's willingness to gamble might depend upon the outcome dimension considered, or on other situation-specific variables [35].

3. Different processes of evaluation. The first two explanations assume that health outcomes are evaluated apart from assessment method, and that an individual's responses compromise between these evaluations and the pursuit of secondary goals such as "using up" the response scale or avoiding gambles. We believe that different assessment methods induce different processes of evaluation. Components of such processes include (a) attention to and perception of stimuli in choice tasks; (b) cognitive activity, such as recalling and taking account of past events and life goals, and selecting reference points, such as expectations or forgone outcomes, against which consequences are evaluated [36]; and (c) the emotional reaction to past, present, and future health states. For example, the higher SG ratings given to intermediate outcomes may result from a negative shift in the perceived value of the worst outcome, surgical mortality. This shift may reflect differences in the psychological processes noted above, such as a more intense emotional reaction to bad outcomes when they are presented in gambles.

Further study of preference scale comparability would benefit from a richer model of the psychological processes underlying evaluations. Rather than conceiving of preferences as fully formed mental entities that wait to be "elicited" from the psyche, we prefer to think that people construct preferences using the tools provided in the value assessment task [23,27]. In this light, divergence among preferences constructed with assessment procedures as different as SG and CS is not surprising.

The practical significance of differences among assessment methods depends upon whether or not they would lead to different choices. Depending on the construction and probabilistic elements of a choice problem, the utility of an alternative can be remarkably insensitive to variability in preference functions for specific outcomes [26]. In our study we substituted CS and SG values for each subject into a simplified decision tree for the coronary artery disease problem (Figure 4). (For this test, the probabilities used represent approximations tailored to the patient described in the workshop materials.) Compared to CS, the relatively more positive evaluation given intermediate outcomes with SG changes the recommended course of action for 36 (60%) of the 60 physicians (Table 4). Thus the method effect does
make a considerable difference for this simplified, didactic decision model.

Since the standard gamble method is based on axiomatic assumptions about preference behavior, it has been advanced as the standard against which other methods should be evaluated [13]. This rationale is challenged, however, by a recent study showing differences between two types of SG techniques, both of which satisfy the same choice axioms [20]. The direct application of these axioms to preference assessment is therefore complicated by the possibility that theoretically equivalent methods may be psychologically distinct. Decision makers who wish to conform to the axioms of utility theory may be at a loss as to which method to choose.

Our study does not resolve debates about whether one method is superior to another. If, as we suggest, different assessment methods lead people to construct different preferences, it may not be valid to consider one method of preference assessment as a standard for another. Our results do demonstrate that clinical decision analysts should expect systematic differences among assessment methods even among people exposed to a thorough introduction to utility assessment. As a practical matter, the sensitivity of clinical decisions to variation in preference scales should be explored carefully.

It should be noted, however, that these issues are not unique to quantitative approaches to clinical decision making. Patients' evaluations of health outcomes are also influenced by more conventional types of interaction with physicians. The pieces of information a physician chooses to communicate or stress can have a major influence on the patient's decision. In fact, a major contribution of preference-scaling methods may lie, not in recommending particular courses of action, but in providing a framework for more thoughtful examination of the factors that already influence clinical decisions.

References

26. von Winterfeldt D, Edwards W: Error in decision analysis. How to create the possibility of large losses by using dominated strategies (Technical Report 75-4). Social Science Research Institute, University of Southern California, 1975
27. Patrick CL, Bush JW, Chen MM: Methods for measuring levels of well-being for a health status index. Health Serv Res 8:228-245, 1975