# Application of Multi-Attribute Utility Theory to Measure Social Preference for Health States

**Article** *in* Operations Research · December 1982

DOI: 10.1287/opre.30.6.1043 · Source: PubMed



**3 authors**, including:

George W Torrance
McMaster University
**196** PUBLICATIONS **26,605** CITATIONS

SEE PROFILE

# Application of Multi-Attribute Utility Theory to Measure Social Preferences for Health States

GEORGE W. TORRANCE, MICHAEL H. BOYLE and SARGENT P. HORWOOD

*McMaster University, Hamilton, Ontario, Canada*

(Received September 1981; accepted July 1982)

A four-attribute health state classification system designed to uniquely categorize the health status of all individuals two years of age and over is presented A social preference function defined over the health state classification system is required. Standard multi-attribute utility theory is investigated for the task, problems are identified and modifications to the standard method are proposed The modified method is field tested in a survey research project involving 112 home interviews. Results are presented and discussed in detail for both the social preference function and the performance of the modified method. A recommended social preference function is presented, complete with a range of uncertainty. The modified method is found to be applicable to the task—no insurmountable difficulties are encountered. Recommendations are presented, based on our experience, for other investigators who may be interested in reapplying the method in other studies.

THE APPLICABILITY of cost-effectiveness analysis as a technique for the evaluation and comparison of alternative health care programs is limited by the difficulty of measuring program effects (outcomes, consequences) in commensurable units across programs of different types. One approach to overcoming this limitation is to determine the ultimate impact of each program on the health states of each individual affected by the program and to use a social preference function defined over the relevant health states as the common unit of measure. (For example, see Fanshel and Bush [1970], Torrance et al. [1972], Torrance [1976a], Weinstein and Stason [1977], Rosser and Watts [1978].)

The general approach in determining such a social preference function is to define a set of health states of interest, to identify a group of subjects (judges), to measure each subject's preferences for the health states, and to aggregate across the subjects to determine the overall social preference function. Within this general approach a number of methods are available for the measurement of the subjects' preferences—these methods have been summarized and compared by Fischer [1979]. Using Fischer's ter-

*Subject classification:* 271 social preferences for multi-attribute health states, 852 measurement of social preferences for health states.



**Operations Research**
Vol. 30, No. 6, November–December 1982

0030-364X/82/3006-1043 $01.25

minology, past work in the measurement of social preferences for health states has generally used *holistic utility assessment methods* (examples include Patrick et al. [1973a, b], Torrance et al. [1973], Torrance [1976b], Rosser and Kind [1978], and Sackett and Torrance [1978]), although there has been one application of a *statistically inferred model* using the *functional measurement* approach (Kaplan et al. [1976]). This paper reports the results of a study which uses a third method—*the explicitly decomposed multi-attribute utility method* using the *conditional utility function-based procedure.* This is, in fact, the classical or standard multi-attribute utility (MAU) method (Keeney and Raiffa [1976], Farquhar [1977]) and, for brevity, it will be referred to as the MAU method. It is an appropriate and relatively efficient method when the health states are described by a multi-attribute classification system. The purpose of this paper is twofold: first, to report the modifications made to the MAU method to adapt it to a survey research project for the measurement of collective social preferences; and second, to report the results of a field study to test the modified MAU method and to develop a specific social preference function.

The paper is divided into five sections. Section 1 contains a description of the multi-attribute health state classification system for which a social preference function is desired. In Section 2 we review the standard MAU method and describe the modifications made to it for this study. Fieldwork and results are presented in Section 3. Section 4 contains a discussion and interpretation of the results, and our conclusions and recommendations are given in Section 5.

## 1. MULTI-ATTRIBUTE HEALTH STATE CLASSIFICATION SYSTEM

The health states in this study are defined according to the four-attribute health state classification system shown in Table I. The system was developed to classify and to follow for life the health outcomes of randomly selected infants in an evaluation of neonatal intensive care (Boyle et al. [1982]). All selected children (age range 2–15 years) have had their current health status classified, their past health pattern reconstructed and their future health pattern forecast using the health state classification system. The children represent a wide variety of disabilities, mostly chronic.

Each attribute in the Health State Classification System is subdivided into a number of levels such that each person can be classified at every point in time into one level on each attribute. A social preference function defined over the health states described by Table I is required. Since each feasible combination of attribute levels defines a unique health state, the system implicitly includes a large number of different states;

TABLE I

HEALTH STATE CLASSIFICATION SYSTEM (AGE ≥ 2 YEARS)

| $X_1$ *PHYSICAL FUNCTION*: MOBILITY AND PHYSICAL ACTIVITY[a] | | |
|---|---|---|
| Level $x_1$ | Code | Description |
| 1 | P1 | Being able to get around the house, yard, neighborhood or community WITHOUT HELP from another person; AND having NO limitation in physical ability to lift, walk, run, jump or bend. |
| 2 | P2 | Being able to get around the house, yard, neighborhood or community WITHOUT HELP from another person; AND having SOME limitations in physical ability to lift, walk, run, jump or bend. |
| 3 | P3 | Being able to get around the house, yard, neighborhood or community WITHOUT HELP from another person; AND NEEDING mechanical aids to walk or get around. |
| 4 | P4 | NEEDING HELP from another person in order to get around the house, yard, neighborhood or community; AND having SOME limitations in physical ability to lift, walk, run, jump or bend |
| 5 | P5 | NEEDING HELP from another person in order to get around the house, yard, neighborhood or community; AND NEEDING mechanical aids to walk or get around. |
| 6 | P6 | NEEDING HELP from another person in order to get around the house, yard, neighborhood or community; AND NOT being able to use or control the arms and legs. |
| $X_2$ *ROLE FUNCTION*: SELF-CARE AND ROLE ACTIVITY[a] | | |
| Level $x_2$ | Code | Description |
| 1 | R1 | Being able to eat, dress, bathe and go to the toilet WITHOUT HELP; AND having NO limitations when playing, going to school, working or in other activities. |
| 2 | R2 | Being able to eat, dress, bathe and go to the toilet WITHOUT HELP; AND having SOME limitations when working, going to school, playing or in other activities. |
| 3 | R3 | Being able to eat, dress, bathe and go to the toilet WITHOUT HELP; AND NOT being able to play, attend school or work. |
| 4 | R4 | NEEDING HELP to eat, dress, bathe or go to the toilet; AND having SOME limitations when working, going to school, playing or in other activities. |
| 5 | R5 | NEEDING HELP to eat, dress, bathe or go to the toilet; AND NOT being able to play, attend school or work. |
| $X_3$ *SOCIAL-EMOTIONAL FUNCTION*: EMOTIONAL WELL-BEING AND SOCIAL ACTIVITY | | |
| Level $x_3$ | Code | Description |
| 1 | S1 | Being happy and relaxed most or all of the time, AND having an average number of friends and contacts with others. |
| 2 | S2 | Being happy and relaxed most or all of the time, AND having very few friends and little contact with others. |
| 3 | S3 | Being anxious or depressed some or a good bit of the time, AND having an average number of friends and contact with others. |
| 4 | S4 | Being anxious or depressed some or a good bit of the time, AND having very few friends and little contact with others. |

*Table I—Continued*

| $X_4$ HEALTH PROBLEM[b] | | |
|---|---|---|
| Level $x_4$ | Code | Description |
| 1 | H1 | Having no health problem. |
| 2 | H2 | Having a minor physical deformity or disfigurement such as scars on the face. |
| 3 | H3 | Needing a hearing aid. |
| 4 | H4 | Having a medical problem which causes pain or discomfort for a few days in a row every two months. |
| 5 | H5 | Needing to go to a special school because of trouble learning or remembering things. |
| 6 | H6 | Having trouble seeing even when wearing glasses. |
| 7 | H7 | Having trouble being understood by others. |
| 8 | H8 | Being blind OR deaf OR not able to speak. |

[a] Multiple choices within each description are applied to individuals as appropriate for their age. For example, a 3-year-old child is not expected to be able to get around the community without help from another person.

[b] Individuals with more than one health problem are classified according to the problem they consider the most serious.

too many to measure preferences explicitly using holistic utility assessment methods. Thus an approach based on MAU theory was selected.

## 2. MULTI-ATTRIBUTE UTILITY THEORY FOR SOCIAL PREFERENCES

### 2.1. MAU Method

Multi-attribute utility (MAU) theory (Keeney and Raiffa, and Farquhar) is concerned with expressing the utilities of multiple-attribute outcomes or consequences as a function of the utilities of each attribute taken singly. The theory specifies several possible functions (additive, multiplicative and multilinear) and the conditions (independence conditions to be met) under which each would be appropriate. As a practical matter Keeney and Raiffa (see p. 298) suggest that for four or more attributes the reasonable models to consider are the additive and the multiplicative. Since our problem contains four attributes, we restrict our attention to these two forms.

Standard MAU theory has been developed for the case of a single decision maker or a single decision making unit. We will first review briefly how this standard theory applies to our multi-attribute problem, and then describe the modifications required to adapt the method for a survey research project to measure collective social preferences. Readers unfamiliar with standard MAU theory who find the remainder of this subsection too condensed may wish to refer to one of the following:

Keeney and Raiffa, Farquhar, or Fischer. All notation used throughout the paper is summarized for easy reference in Table II.

The conventional method for measuring utilities is the standard gamble technique, a lottery-based procedure first proposed by von Neumann and

TABLE II

NOTATION

$X_j$, $j$ = 1, 2, 3, 4 represents attribute $j$ (Table I)

$x_j$ represents the level on attribute $j$ (Table I)

$x_j^*$ represents the best (most preferred) level on attribute $j$ according to the individual

$x_j^o$ represents the worst (least preferred) level on attribute $j$ according to the individual

$x_j^a$ represents a prespecified good level (the level most preferred by most people) on attribute $j$

$x_j^b$ represents a prespecified bad level (the level least preferred by most people) on attribute $j$

$v_j^*(x_j)$ is the individual's single-attribute value function for attribute $j$, on the value scale with $v_j^*(x_j^*) = 1$ and $v_j^*(x_j^o) = 0$

$v_j^a(x_j)$ is the individual's single-attribute value function for attribute $j$, on the value scale with $v_j^a(x_j^a) = 1$ and $v_j^a(x_j^b) = 0$

$v_G$ is the group single-attribute value function (the arithmetic mean of the individual values)

$\bar{v} = 1 - v$ is the associated disvalue function

$u_j^*(x_j)$, $u_j^a(x_j)$, $u_G$ and $\bar{u}$ have parallel definitions but for utility rather than value

$\underline{x} = (x_1, x_2, x_3, x_4)$ represents a multi-attribute health state with the specified levels on each attribute

$\underline{x}^*$, $\underline{x}^o$, $\underline{x}^a$, $\underline{x}^b$, $\underline{x}'$ represent specific multi-attribute health states formed according to the rule $\underline{x}^k = (x_1^k, x_2^k, x_3^k, x_4^k)$, where $k$ can be *, 0, $a$, $\cdots$

$\underline{x}^h$ represents the specific health state, full healthy life

$\underline{x}^d$ represents the specific health state, death at birth

$(x_j^k, x_{\bar{j}}^l)$ represent specific multi-attribute health states formed according to the rule that attribute $j$ is at level $x^k$ and all other attributes are at level $x^l$

$u^*(\underline{x})$ is the individual's multi-attribute utility function, on the utility scale with $u^*(\underline{x}^*) = 1$ and $u^*(\underline{x}^o) = 0$

$u^a(\underline{x})$ is the individual's multi-attribute utility function, on the utility scale with $u^a(\underline{x}^a) = 1$ and $u^a(\underline{x}^b) = 0$

$u^h(\underline{x})$ is the individual's multi-attribute utility function, on the utility scale with $u^h(\underline{x}^h) = 1$ and $u^h(\underline{x}^d) = 0$

$\bar{u} = 1 - u$ is the associated disutility function

$U_1(\underline{x})$ is the group multi-attribute utility function on the utility scale with $U_1(\underline{x}^h) = 1$ and $U_1(\underline{x}^d) = 0$, formed by aggregating at the final level. i.e. $U_1(\underline{x})$ is the mean $u^h(\underline{x})$

$U_2(\underline{x})$ is the group multi-attribute utility function on the utility scale with $U_2(\underline{x}^h) = 1$ and $U_2(\underline{x}^d) = 0$, formed by aggregating at the basic level (this is defined more precisely later)

$L = \langle p_1\underline{x}', p_2\underline{x}'' \rangle$ is a lottery (an uncertain event) with probability $p_1$ that the outcome will be $\underline{x}'$ and probability $p_2$ that it will be $\underline{x}''$

Morgenstern [1953]. In this method two reference outcomes are established, for example $\underline{x}^*$ and $\underline{x}^0$, and the utility of the other outcomes are measured relative to these two. For an intermediate outcome $\underline{x}$ the subject is asked to determine the probability $p$ such that s/he is indiffer-

ent between the lottery $\langle p\underline{x}^*, (1-p)\underline{x}^0\rangle$ and $\underline{x}$ for certain. Then $u(\underline{x}) = p$ on the utility scale where $u(\underline{x}^*) = 1$ and $u(\underline{x}^0) = 0$.

*Additive Form*

Additive independence exists if each attribute is additive independent of the other attributes. Attribute $j$ is additive independent of attributes $\bar{j}$ if the subject is indifferent between lottery $L_1$ and lottery $L_2$ for all values of $x_j'$ and $x_j''$, where $L_1 = \langle 0.5\underline{x}', 0.5\underline{x}''\rangle$ and $L_2 = \langle 0.5(x_j', x_j''), 0.5(x_j'', x_j')\rangle$.

If additive independence exists, the multi-attribute utility function is additive.

$$u^*(\underline{x}) = \sum_{j=1}^{4} k_j u_j^*(x_j) \tag{1}$$

where

$$\sum_{j=1}^{4} k_j = 1 \tag{2}$$

and

$$k_j = u^*(x_j^*, x_j^0), \qquad j = 1, 2, 3, 4. \tag{3}$$

One simple method to determine the function is to measure each single-attribute utility function $u_j^*(x_j)$ separately relative to $u_j^*(x_j^*) = 1$ and $u_j^*(x_j^0) = 0$, while holding the levels of all other attributes constant. Then the $k$'s are determined by measuring the utility of the specific multi-attribute health states in (3) relative to $u^*(\underline{x}^*) = 1$ and $u^*(\underline{x}^0) = 0$.

Alternatively and equivalently, one can define and measure the additive multi-attribute disutility function

$$\bar{u}^*(\underline{x}) = \sum_{j=1}^{4} c_j \bar{u}_j^*(x_j) \tag{4}$$

where

$$\sum_{j=1}^{4} c_j = 1 \tag{5}$$

and

$$c_j = \bar{u}^*(x_j^0, x_j^*), \qquad j = 1, 2, 3, 4. \tag{6}$$

Again, the function can be determined by measuring each single-attribute disutility function $\bar{u}_j^*(x_j)$ separately, relative to $\bar{u}_j^*(x_j^*) = 0$ and $\bar{u}_j^*(x_j^0) = 1$, while holding the levels of all other attributes constant. The $c_j$ can be determined from (6) by measuring the disutility of the specified multi-attribute health states relative to $\bar{u}^*(\underline{x}^*) = 0$ and $\bar{u}^*(\underline{x}^0) = 1$. Note that for the additive case $c_j = k_j$ all $j$, but the $c_j$ notation is introduced here for convenience later.

*Multiplicative Form*

Mutual utility independence exists if every subset of $\{X_1, X_2 X_3, X_4\}$ is utility independent of its complement. A subset of attributes is utility independent of its complementary set if the conditional preference order for lotteries involving only changes in the levels of attributes in the subset does not depend on the levels at which the attributes in the complementary set are held fixed. Additive independence implies mutual utility independence, but not vice versa.

If mutual utility independence exists, the multi-attribute utility function is additive, or multiplicative of the form

$$u^*(\underset{\sim}{x}) = (1/k)[\prod_{j=1}^{4} (1 + kk_j u_j^*(x_j)) - 1]. \tag{7}$$

One method to determine the function is to measure each $u_j^*(x_j)$, determine the four $k_j$ values from (3), and find the $k$ value by iteratively solving (8), which is (7) for $\underset{\sim}{x} = \underset{\sim}{x}^*$.

$$1 + k = \prod_{j=1}^{4} (1 + kk_j). \tag{8}$$

Parameter $k$ is related to parameters $k_j$ as follows:

$$\text{if } \sum_{j=1}^{4} k_j > 1, \text{ then } -1 < k < 0, \tag{9a}$$

$$\text{if } \sum_{j=1}^{4} k_j = 1, \text{ then } k = 0 \text{ and the additive model holds, and} \tag{9b}$$

$$\text{if } \sum_{j=1}^{4} k_j < 1, \text{ then } k > 0. \tag{9c}$$

The three cases can be distinguished in terms of the multivariate risk posture which they represent (Richard [1975]). Case (9a) represents multivariate risk aversion, case (9b) multivariate risk neutrality and case (9c) multivariate risk seeking behavior. The attributes in case (9a) can be characterized as "substitutes" while those in (9c) are "complements" (Keeney and Raiffa, p. 240). The intuitive interpretation of this characterization is that substitute attributes are such that an improvement in one is relatively satisfying, while an improvement on two or more is not that much better. Conversely, with complementary attributes an improvement on any one alone is not very useful, while a simultaneous improvement on several is much better.

As an alternative to (7) one can define and measure the multiplicative multi-attribute disutility function

$$\bar{u}^*(\underset{\sim}{x}) = (1/c)[\prod_{j=1}^{4} (1 + cc_j \bar{u}_j^*(x_j)) - 1] \tag{10}$$

where

$$(1 + c) = \prod_{j=1}^{4} (1 + cc_j) \tag{11}$$

and

$$\text{if } \sum_{j=1}^{4} c_j > 1, \text{ then } -1 < c < 0, \tag{12a}$$

if $\sum_{j=1}^{4} c_j = 1$, then $c = 0$ and the additive model holds, and (12b)

if $\sum_{j=1}^{4} c_j < 1$, then $c > 0$. (12c)

The function can be found by measuring the $\bar{u}_j^*(x_j)$, determining the $c_j$ from (6) and $c$ from (11).

Case (12a) represents multivariate risk seeking behavior, case (12b) multivariate risk neutrality and case (12c) multivariate risk aversion. In case (12a) the disattributes can be characterized as substitutes, while in case (12c) they are complements. The intuitive notion here is that in case (12a) a reduction on any one attribute is bad, while a reduction on two or more is not that much worse. Conversely, in case (12c) a reduction on any one attribute alone is not so bad, while a simultaneous reduction on several is very serious.

The utility and disutility formulations are related to each other in the following ways. There are three possibilities.

*Case 1.* $k < 0$, $c > 0$, multiplicative model, multivariate risk aversion, attributes are substitutes, disattributes are complements.

*Case 2.* $k = 0$, $c = 0$, additive model, multivariate risk neutrality, attributes have no preference interaction.

*Case 3.* $k > 0$, $c < 0$, multiplicative model, multivariate risk seeking, attributes are complements, disattributes are substitutes.

In health applications the disutility formulation (10) appears to be more natural and better suited than the utility formulation (7). This occurs because most subjects, at the time of the interview, are at or near health state $\underset{\sim}{x}^*$ and can, therefore, relate more readily to the health states in (6) than to those in (3). In (6) the subjects only need to consider changes from their current state one attribute at a time, whereas in (3) they must deal with simultaneous changes in three attributes.

*Procedure*

Classical multi-attribute utility theory, as described above, could be applied to our problem using the following four steps on each subject. *Step 1:* Establish and verify the necessary independence conditions. *Step 2:* Measure the single-attribute disutility functions $\bar{u}_j^*(x_j)$. *Step 3:* Measure the disutility of the multi-attribute "corner" states in (6) to determine the $c_j$ parameters. *Step 4:* Use (5) or (11) as appropriate to complete the determination of the parameters for the additive (4) or the multiplicative (10) form respectively. The result of these four steps would be a multi-attribute disutility function $\bar{u}^*(\underset{\sim}{x})$ for each subject scaled such that $\bar{u}^*(\underset{\sim}{x}^*) = 0$ and $\bar{u}^*(\underset{\sim}{x}^0) = 1$. However, several problems emerge in attempting to translate this method to a survey research project designed to measure collective social preferences. The problems and the modifications we adopted to overcome them are described in the next subsection.

## 2.2. Modifications to MAU Method

*Independence Condition*

The establishment and verification of the independence conditions is normally a tedious, exacting and time-consuming task requiring extensive interviewer-subject interaction (see, for example, Keeney and Raiffa). This approach is only feasible for studies with a small number of subjects. As an alternative to Step 1 we elected to assume the existence of mutual utility independence (note that this assumption is fully consistent with our previous action of restricting the investigation to the additive and the multiplicative MAU models) and to test this assumption later with data obtained by measuring the disutility of additional multi-attribute health states. These additional "test" states are not used in the construction of the multi-attribute disutility functions, but are used later to test their fit.

*Measurement Techniques*

A second problem for a survey research project with many subjects selected randomly from the general public is the complexity and difficulty of administering the usual lottery-based utility measurement techniques like the von Neumann-Morgenstern standard gamble (Torrance [1976b], Wolfson et al. [1982]). We circumvented this problem in our study by using simpler measurement methods not involving probabilities. All our single-attribute measurements were made using a category scaling method, while our multi-attribute measurements used the time trade-off procedure.

The category scaling method used a visual analog device called a "feeling thermometer." It is a thermometer-shaped 0–100 scale on a felt board with 0 labeled "least desirable" and 100 labeled "most desirable." The "levels" of the attribute being measured are printed on narrow foam sticks pointed at each end which can be placed on the felt board beside the thermometer. The subject was asked to imagine being in these situations for a lifetime with everything else normal, or average. S/he was asked to place the most desirable level $x_j^*$ at 100, the least desirable $x_j^0$ at 0, and the others in between in order of desirability, with ties allowed, and spaced such that the relative distance *between* the levels corresponds to her/his feelings about the relative *differences* in desirability. Then for any level $x_j'$ the value of the level, $v_j^*(x_j')$, on the value scale where $v_j^*(x_j^*) = 1$ and $v_j^*(x_j^0) = 0$ is the thermometer reading beside the level divided by 100. Later in the interview, after considerable intervening material the subject was asked to remeasure one of the attributes (randomly determined) as a reliability check.

Because uncertainty is not used in the category scaling method, it measures a value function $v_j^*(x_j)$ as opposed to a utility function $u_j^*(x_j)$

(Keeney and Raiffa). To convert these values to utilities we need to know the relationship, for health state preferences, between value functions as measured by category scaling and utility functions as measured by a lottery technique like the von Neumann-Morgenstern standard gamble. Two studies have investigated this relationship. Our previous work (Note 1) found for population means the following power curve relationship between disvalue and disutility

$$\bar{u} = \bar{v}^{1.6} \tag{13}$$

while recent work by Wolfson and his colleagues (Note 2) provides independent confirmation of the general nature of this relationship although not of the specific parameter value. In the work reported here we have used (13) to convert the single-attribute measures from values to utilities. It should be noted that (13) represents a particular attitude toward uncertainty. By adopting (13) we do not measure the subjects' uncertainty attitudes, but assume they are the same as those of previous subjects. Later, as part of the sensitivity analysis, we investigate the impact of this assumption.

The time trade-off procedure as developed by Torrance et al. [1972] and later modified to handle states worse than death (Torrance [1982]) was used for all the multi-attribute measurements. For chronic state $\underset{\sim}{x}$ preferred to death the subject was asked to determine the time $t$ such that s/he is indifferent between (i) a lifetime (70 years) in the chronic state and (ii) a healthy but shorter life of $t$ years. Then $v^h(\underset{\sim}{x}) = t/70$ on the value scale where $v^h(\underset{\sim}{x}^h) = 1$ and $v^h(\underset{\sim}{x}^d) = 0$. For chronic state $\underset{\sim}{x}$ dispreferred to death the subject was asked to determine the time $t$ such that s/he is indifferent between (i) a healthy life for $t$ years followed by the remainder of life (to age 70) in $\underset{\sim}{x}$ and (ii) to die in hospital shortly after birth. Then $v^h(\underset{\sim}{x}) = \alpha t/(t - 70)$ where $\alpha$ is a parameter of the instrument and the interviewing procedure such that the lower limit of the instrument is $-1.0$. This maintains symmetry between the value scale for states preferred to death which runs from 0 to 1, and the scale for states dispreferred to death which runs from 0 to $-1$. In all cases the time trade-off procedure was supplemented with visual aids. Later in the interview, after intervening material, the subject was asked to remeasure one of the multi-attribute health states using the time trade-off method as a reliability check.

Since lotteries are not used in the time trade-off procedure, the results do not incorporate the subject's attitude toward uncertainty. However, previous work (Torrance et al. [1973], Torrance [1976b]) demonstrated that for states preferred to death, population means measured by this method are empirically equivalent to those obtained from a lottery-based technique, the von Neumann-Morgenstern standard gamble, which does

incorporate attitude toward uncertainty. Consequently, in this study the time trade-off results are assumed to represent utilities. For states dispreferred to death no such previous work exists and, in the absence of any, the scaled (to $-1.0$) results of the time trade-off technique are tentatively accepted as utilities.

*Extreme Levels*

A problem can arise in measuring the disutility of the "corner" states in (6) if the subjects differ in the levels which they rate as best and worst on the individual attributes. The corner states for a subject are defined by the best and worst levels as perceived by that particular subject. This can complicate the interview since the corner states, therefore, are not known until the interview is in progress and the corner states may differ from subject to subject. One solution is to cope with the situation during the interview by designing a flexible interviewing format. Another approach is to ensure that each attribute contains one extremely good and one extremely bad level, so that most if not all subjects would agree on the best and worst. A variation on this approach used by Krischer [1976] is to prespecify to the subject which is the best and which is the worst level, and ask the subject to rate the others relative to these two extremes. A third alternative (which we used) is to prespecify for each attribute a good level, $x_j^a$, and a bad level, $x_j^b$, but not tell the subject. In this way the subject is unconstrained in providing her/his true preferences; rather the necessary corrections are handled later in the calculations. The prespecified levels, $x_j^a$ and $x_j^b$, are used in (6) to create a set of common corner states for measurement on all subjects, as follows:

$$c_j = \bar{u}^a(x_j^b, x_j^a), \qquad j = 1, 2, 3, 4. \tag{14}$$

Then, for those subjects for whom the measured states are not true corner states, the calculation method requires a slight modification. In Step 2 before (13) is applied the single-attribute disvalue functions $\bar{v}_j^*(x_j)$, are transformed to the $\bar{v}_j^a(x_j)$ scale by the following positive linear transformation

$$\bar{v}_j^a(x_j) = [\bar{v}_j^*(x_j) - \bar{v}_j^*(x_j^a)]/[\bar{v}_j^*(x_j^b) - \bar{v}_j^*(x_j^a)] \tag{15}$$

where $\bar{v}_j^*(x_j^b) > \bar{v}_j^*(x_j^a)$ relationship. The remainder of the method is unaffected except that the resulting multi-attribute disutility function is $\bar{u}^a(\underline{x})$ as oposed to $\bar{u}^*(\underline{x})$.

*Social Aggregation*

A significant issue in applying multi-attribute utility theory to measure collective social preferences is the question of how one should aggregate individual preferences into social preferences. For example, the methods

described above will produce individual multi-attribute disutility functions $\bar{u}^a(\underline{x})$. The question of aggregating such individual cardinal preferences into collective social preferences has been addressed by a number of authors, and different sets of assumptions lead to different results—for example, Hildreth [1953], Harsanyi [1975], and Keeney [1976] all argue that aggregation is valid while Kalai and Schmeidler [1977] argue the converse. We agree with those who favor aggregation, and we underscore the point made by Harsanyi, and Dyer and Sarin [1979] that such comparisons of individual preferences are common practice—indeed, in order to make social decisions, and in the very process of making those decisions, individual preferences *must* be and *are* compared. The question, then, is not whether to make such comparisons but how to make them. In the spirit of Hildreth we establish two clearly defined outcomes, one good and one bad, as anchor points (but not necessarily end points) for the utility scale. The good outcome is a normal healthy life (defined as birth to age 70), $\underline{x}^h$, and is given a utility of 1 for each individual. The bad outcome is death in hospital shortly after birth, $\underline{x}^d$, and is given a utility of 0 for each individual. The central basis for the aggregation is that the difference in utility between these two outcomes is set equal across people. The aggregation method is the arithmetic mean. The mean is the method recommended by Harsanyi, and is also the method we obtain from the models of Hildreth, or Keeney by assigning each individual equal weight. The overall method is also consistent with the recommended practice in health program evaluation in which for each individual, regardless of who s/he is, immediate death contributes zero and one healthy year contributes one unit to the effectiveness of the program (Weinstein and Stason). To implement the method we measure, for each individual, the utility of $\underline{x}^a$ and $\underline{x}^b$ relative to $u^h(\underline{x}^d) = 0$ and $u^h(\underline{x}^h) = 1$ and transform the individual's multi-attribute disutility function to the $u^h$ utility scale by the negative linear transformation

$$u^h(\underline{x}) = u^h(\underline{x}^a) - [u^h(\underline{x}^a) - u^h(\underline{x}^b)]\bar{u}^a(\underline{x}). \tag{16}$$

If $u_i^h(\underline{x})$ is the resulting utility function for individual $i$ of $m$ individuals, the social utility function is simply

$$U_1(\underline{x}) = (1/m) \sum_{i=1}^{m} u_i^h(\underline{x}) \tag{17}$$

where the reference states, healthy and dead, have values 1 and 0 respectively; that is, $U_1(\underline{x}^h) = 1$, $U_1(\underline{x}^d) = 0$.

$U_1(\underline{x})$ is obtained by performing the social aggregation at the final level of the process after each individual's unique $u^h(\underline{x})$ has been determined. An alternative approach is to perform the social aggregation at the basic level on the measured data and develop a group solution all the way through the process. This alternative approach is introduced as a con-

venient approximation although it does not have the theoretical underpinning of the first approach. The method simply consists of measuring the fundamental utilities on a common scale right from the beginning and using the arithmetic mean as the group measure. Specifically, the single-attribute disvalue functions for each individual $\bar{v}_j^*(x_j)$ as measured by the category scaling method are first converted to the common $\bar{v}^a$ scale using (15), and are then aggregated using the arithmetic mean into the group functions $\bar{v}_{Gj}^a(x_j)$. These group disvalue functions $\bar{v}_{Gj}^a(x_j)$ are converted to group disutility functions $\bar{u}_{Gj}^a(x_j)$ by (13). The individual utilities for the corner states in (14) and for states $\underline{x}^a$ and $\underline{x}^b$ are measured by the time trade-off method on the $u^h$ utility scale, and the means are used to represent the group preferences $u_G{}^h(\underline{x})$. These are converted to disutilities on the $\bar{u}^a$ scale using

$$\bar{u}_G{}^a(\underline{x}) = [u_G{}^h(\underline{x}^a) - u_G{}^h(\underline{x})]/[u_G{}^h(\underline{x}^a) - u_G{}^h(\underline{x}^b)] \qquad (18)$$

which is the inverse of (16) at the group level, and the disutilities $\bar{u}_G{}^a(\underline{x})$ of the four corner states in (14) are the $c_j$ values for the group. The group disutility function $\bar{u}_G{}^a(\underline{x})$ is specified using (4) and (5) or (10) and (11) as appropriate. Finally, (16) is used at the group level to specify the desired social utility function as follows:

$$U_2(\bar{x}) = u_G{}^h(\underline{x}^a) - [u_G{}^h(\underline{x}^a) - u_G{}^h(\underline{x}^b)]\bar{u}_G{}^a(\underline{x}) \qquad (19)$$

where again the reference states of healthy and dead have values 1 and 0 respectively; that is, $U_2(\underline{x}^h) = 1$, $U_2(\underline{x}^d) = 0$.

### 2.3. Summary of Method

To summarize, the method we used was to measure individual single-attribute value functions $v_j^*$ using the category scaling method and individual utilities $u^h$ for multi-attribute states using the time trade-off technique. The $v_j^*$ values were converted to $\bar{v}_j^*$ disvalues using $\bar{v}_j^* = 1 - v_j^*$ and then to $\bar{v}_j^a$ disvalues using (15). For model $U_1$ the individual $\bar{v}_j^a$ disvalues are converted to $\bar{u}_j^a$ disutilities using (13), the individual $u^h$ utilities are converted to $\bar{u}^a$ disutilities using the inverse of (16), the $c_j$ values are found from (14), (4) and (5) or (10) and (11) as approporiate are used to determine the individual $\bar{u}^a(\underline{x})$ functions, and (16) and (17) are used to determine the social preference function. For model $U_2$ the individual $\bar{v}_j^a$ disvalues and the individual $u^h$ utilities are aggregated using the arithmetic mean before being converted by (13) and the inverse of (16) respectively, the $c_j$ values are found from (14), (4) and (5) or (10) and (11) are used to determine the group $\bar{u}_G{}^a(\underline{x})$ function, and (16) is used to specify the social preference function.

## 3. FIELD MEASUREMENTS AND RESULTS

*Fieldwork*

Parents of school children were used as subjects. A random sample of 148 names was drawn by the Research Department of the Hamilton Board of Education. Interviews were conducted in the subject's home by the professional interviewing staff of a local survey research firm. (Copies of the interviewer's manual and instructions are available upon request.) For that part of the interview reported here (chronic states) subjects were told to imagine that they are in the health situation described and that it lasts for a lifetime (birth to age 70).

Each subject provided single-attribute value functions $v_j^*(x_j)$ for each attribute using the category scaling method (feeling thermometer), and utilities $u^h(\underset{\sim}{x})$ for the seven multi-attribute health states shown in Table IV using the time trade-off method. States A–E in Table IV are used in the construction of the multi-attribute utility function, while states F and G are used later to test the fit. States A–D are the four "corner" states required in (14) to determine the $c_j$ values, while state E is $\underset{\sim}{x}^b$ required in (16) to convert between $\bar{u}^a$ and $u^h$. The states in Table IV were presented to the subjects on printed cards containing the descriptions of the attribute levels shown in the table. The subjects were asked to interpret the cards as follows: "Each situation lasts a lifetime, let's say all the way from birth to age 70. Please think of yourself as being in these situations but in every other way as healthy as possible."

Completed interviews were obtained from 76% (112/148) of the names drawn in the random sample. The remaining 24% can be divided into those who could not be contacted (1%), those contacted but ineligible due to a language barrier (8%), those who refused to participate (14%) and those who broke off the interview (1%).

Seventy-eight percent (87/112) of completed interviews produced usable data. In the remaining cases the subject gave at least one response that indicated confusion with regard to the measurement task. On the single-attribute measurement task a respondent was categorized as "confused" if s/he ranked any level as more desirable than the prespecified level 1. On the multi-attribute measurement task a respondent was categorized as "confused" if s/he ranked any state as less desirable than state E. Ten respondents displayed confusion on the one-attribute task, 11 on the multi-attribute task and 4 on both. Confused responders were eliminated from the data; all results are calculated from the remaining 87 cases.

In measuring the single-attribute functions, the modification involving $x_j^a$ and $x_j^b$ was used. Based on a pilot study $x_j^a$ was established as P1, R1, S1 and H1 and $x_j^b$ was P6, R5, S4 and H8 for the four attributes

respectively. These were the levels most frequently designated as best and worst in the pilot study. All 87 respondents selected each $x_j^a$ as the best, i.e. $x_j^* = x_j^a$, all $j$. (Recall, those who did not were ruled out as "confused.") On the other hand, not all agreed that $x_j^b$ was the worst on each attribute. Twenty-one of the 87 subjects (24%) disagreed with one or more $x_j^b$, 16 of these disagreed on only one attribute while 5 disagreed on 2 attributes. Thus there were 26 disagreements out of a total of 348 $x_j^b$'s (7.5%). The disagreements were primarily in attribute 4 as would be expected, the distribution of disagreements by attribute was (5, 3, 4, 14). Since $x_j^* = x_j^a$, all $j$, for all respondents, $\bar{v}_j^*(x_j^a) = 0$ and (15) simplifies to

$$\bar{v}_j^a(x_j) = \bar{v}_j^*(x_j)/\bar{v}_j^*(x_j^b). \tag{20}$$

Note the simplicity of (20) in that it can be applied to *all* the measurements and where it is not required it will be made inoperative automatically because in that case $\bar{v}_j^*(x_j^b) = 1$.

*Results*

Four single-attribute disvalue functions were determined for each of the 87 subjects using transformation (20). These were converted to value functions and are reported in aggregate form in Table III.

The utility to each individual for each of the seven multi-attribute health states was measured directly on the $u^h$ utility scale with the time trade-off technique. The findings are reported in aggregate form in Table IV.

The reliability of the measurement methods is reported in Table V. Reliability is based on replicated measures taken during the same interview. The correlation coefficient $r$ is the product moment (Pearson's) correlation coefficient between the original measure and the repeated measure. The precision $\sigma_e$ is the standard error of measurement and is calculated from $\sigma_e = \sqrt{d^2/2N}$, where $d$ is the difference between the original measure and the repeated measure.

For the case of social aggregation at the final level, leading to $U_1(\underline{x})$, each individual's multi-attribute disutility function $\bar{u}^a(\underline{x})$ must be determined. As part of this determination, the $c_j$ parameters from (14) and, if appropriate, the $c$ parameter from (11) must be calculated. A frequency distribution of the resulting parameters is given in Table VI.

For the additive model to be appropriate an individual's four $c_j$ values should sum to 1.0 yielding a $c$ value of 0. As can be seen from Table VI this was not the case for any of the subjects. In fact, very few $c$ values were anywhere close to 0; the closest being 0.11 and this was the only one that fell in the interval $0 \pm 0.25$; while only 7 fell in the interval $0 \pm 0.50$. Thus, the multiplicative model was selected as the more appropriate

TABLE III
SINGLE-ATTRIBUTE VALUE FUNCTIONS ($N = 87$)

| Physical Function | | | Role Function | | | Social Emotional Function | | | Health Problem | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Level | Mean value | Standard error | Level | Mean value | Standard error | Level | Mean value | Standard error | Level | Mean value | Standard error |
| P1 | 1.00 | 0.000 | R1 | 1.00 | 0.000 | S1 | 1.00 | 0.000 | H1 | 1.00 | 0.000 |
| P2 | 0.62 | 0.082 | R2 | 0.71 | 0.021 | S2 | 0.65 | 0.027 | H2 | 0.49 | 0.040 |
| P3 | 0.38 | 0.101 | R3 | 0.32 | 0.019 | S3 | 0.25 | 0.026 | H3 | 0.47 | 0.047 |
| P4 | 0.37 | 0.021 | R4 | 0.30 | 0.022 | S4 | 0.00 | 0.000 | H4 | 0.46 | 0.037 |
| P5 | 0.10 | 0.085 | R5 | 0.00 | 0.000 | | | | H5 | 0.30 | 0.062 |
| P6 | 0.00 | 0.000 | | | | | | | H6 | 0.25 | 0.054 |
| | | | | | | | | | H7 | 0.22 | 0.074 |
| | | | | | | | | | H8 | 0.00 | 0.000 |

TABLE IV

UTILITY OF MULTI-ATTRIBUTE HEALTH STATES ($N = 87$)

| State | Attribute Levels Described on Card | Percentage of Times Rated Worse than Death | Mean Utility | Standard Error |
|---|---|---|---|---|
| $\underline{x}^h$ | No card—healthy | — | 1.00 | 0.000 |
| A | (P6 ) | 61 | 0.01 | 0.069 |
| B | ( R5 ) | 59 | −0.01 | 0.063 |
| C | ( S4 ) | 28 | 0.45 | 0.053 |
| D | ( H8) | 26 | 0.39 | 0.065 |
| E | (P6, R5, S4, H8) | 80 | −0.39 | 0.064 |
| F | (P2, R2, H4) | 13 | 0.67 | 0.037 |
| G | (P5, R2, H5) | 31 | 0.31 | 0.064 |
| $\underline{x}^d$ | No card—death | — | 0.00 | 0.000 |

model for all subjects. Each individual's multiplicative multi-attribute disutility function was used in (16) and the results of this in (17) to give the desired social utility function $U_1(\underline{x})$. The function is not reported here as it is only available in tabular form and the table has 960 entries. However, copies are available upon request.

The alternative approach to social aggregation leads to the function $U_2(\underline{x})$. The group single-attribute value functions $v_{Gj}^{\alpha}(x_j)$ are given in Table III, "Mean Value" column. The group utilities for the multi-attribute states $u_G{}^h(\underline{x})$ are shown in Table IV, "Mean Utility" column. Four of these, A–D, are transformed to disutilities using (18) with $u_G{}^h(\underline{x}^a) = 1$ because in our study $\underline{x}^a = \underline{x}^h$ and with $u_G{}^h(\underline{x}^b) = -0.39$ because $\underline{x}^b = E$. It follows from (14) that these disutilities are the group $c_j$ parameters; the values are $(c_1, c_2, c_3, c_4) = (0.72, 0.73, 0.40, 0.44)$. Since $\sum_{j=1}^{4} c_j > 1$, we know from (12a) that $-1 < c < 0$, and solving (11) gives

TABLE V

RELIABILITY OF MEASURES

| Measure | Scale Length $L$ | Replicated Measures $N$ | Correlation Coefficient $r$ | Precision $\sigma_e$ | Precision as a Proportion of Scale Length $\sigma_e/L$ |
|---|---|---|---|---|---|
| Single-attribute measures (category scaling): | | | | | |
| Physical function | 0, 1 | 132 | 0.93 | 0.095 | 0.095 |
| Role function | 0, 1 | 105 | 0.94 | 0.093 | 0.093 |
| Social-emotional function | 0, 1 | 84 | 0.86 | 0.153 | 0.153 |
| Health problem | 0, 1 | 184 | 0 87 | 0.122 | 0.122 |
| Multi-attribute measures (time trade-off): | | | | | |
| Chronic states | −1, 1 | 87 | 0.88 | 0.240 | 0.120 |

$c = -0.97$. Thus the $U_2$ social preference function developed in this study can be specified in total as follows:

$$\bar{u}^a_{Gj}(x_j) = [1 - v^a_{Gj}(x_j)]^{1.6} \tag{21a}$$

$$\bar{u}_G{}^a(\underline{x}) = (1/c)[\prod_{J=1}^{4} (1 + cc_j\bar{u}^a_{Gj}(x_j)) - 1] \tag{21b}$$

$$U_2(\underline{x}) = 1 - 1.39\bar{u}_G{}^a(\underline{x}). \tag{21c}$$

The complete $U_2$ social preference function for each of the 960 health states in Table I can be determined using equation set (21) in conjunction with the data in Table III and the $c_j$ and $c$ values specified above.

How do the results compare between the two methods of aggregation? Utility values were calculated for each of the 960 states using each method of aggregation, $U_1(\underline{x})$ and $U_2(\underline{x})$, and the two sets of data were compared. The two sets of utilities are highly correlated ($r = 0.995$), however, there is a small but statistically significant ($p < 0.0001$) bias:

TABLE VI

PARAMETERS FOR $\bar{u}^a(\underline{x})$

| Parameter | $c_1$ | $c_2$ | $c_3$ | $c_4$ | Parameter | $c$ |
|---|---|---|---|---|---|---|
| Frequency distribution: | | | | | Frequency distribution: | |
| $c_j = 1$ | 29 | 29 | 6 | 15 | | |
| $0.5 \leq c_j < 1$ | 36 | 43 | 27 | 19 | $c > 0$ | 8 |
| $0 < c_j < 0.5$ | 22 | 15 | 54 | 53 | $c = 0$ | 0 |
| $c_j = 0$ | 0 | 0 | 0 | 0 | $c < 0$ | 79 |
| Total frequency | 87 | 87 | 87 | 87 | Total frequency | 87 |
| Median | 0.81 | 0.78 | 0.30 | 0.30 | Median | −1.00 |

$U_1(\underline{x}) - U_2(\underline{x})$ has a range of −0.09 to 0.09 with a mean value of 0.01. An analysis of the differences shows no particular systematic pattern.

How well does the model fit the two test points F and G? The mean measured utility for these two test points is 0.67 and 0.31 (Table IV), while the model results for the same states are 0.50 and 0.01.

## 4. DISCUSSION

### *Fieldwork*

Respondent eligibility and participation rates are consistent with our previous work in the field (Torrance [1976b], Sackett and Torrance). The current study reconfirms our former findings that a high proportion of eligible subjects will participate in these studies and that very few participants will break off the interview. We take this as evidence that the general thrust of the study as well as the specific measurement tasks are found to be acceptable by the general public.

On the other hand, the level of respondent confusion is disappointing,

although it is consistent with other studies in which interviewer intervention was prohibited. For example, Krischer (see his Table I) reports even higher levels of respondent confusion in a study using self-administered questionnaires; while Kaplan et al. [1979] noted lower, but substantial, confusion rates (17%) in a study using structured interviews. To ameliorate this problem in future work we recommended that interviewer intervention be allowed in clearly defined cases of respondent confusion like those in our study. The interviewer's role would be to identify and explain the apparent inconsistency to the subject but not to insist on its rectification. In addition respondent confusion on the single-attribute task could be reduced by avoiding attributes with a double content like our attributes $X_1$, $X_2$ and $X_3$ (Table I). It is instructive to note that all of our single-attribute respondent confusion came from these attributes, and none came from attribute $X_4$ which contains only one construct. It would appear that the double content of attributes $X_1$, $X_2$ and $X_3$ overloaded the information processing capability of some of the respondents.

*Extreme Levels*

The technique of measuring a common set of corner points involving $x_j^a$ and $x_j^b$ on *all* respondents regardless of their particular $x_j^*$ or $x_j^0$ preferences proved useful in practice, but what are the full implications of doing this? Is it always suitable? Does it require the satisfaction of additional assumptions regarding the underlying utility structure? Does it introduce additional measurement error? First, the method as described in this paper is not suitable for any individual who prefers $x_j^b$ to $x_j^a$, on any attribute; fortunately, this did not happen in our study. Second, the method can be viewed as simply determining the subject's multi-attribute disutility function $\bar{u}^*(\underline{x})$ on the reduced multi-attribute space obtained by omitting on each attribute any $x_j$ level preferred to $x_j^a$ and any $x_j$ level dispreferred to $x_j^b$. Given these omissions, $x_j^a = x_j^*$ and $x_j^b = x_j^0$, on each attributes, and $\bar{u}^*(\underline{x}) = \bar{u}^a(\underline{x})$ can be found in the conventional way. Then, assuming that the independence conditions necessary for $\bar{u}^*(\underline{x})$ on the restricted space also apply over the enlarged space, the $\bar{u}^a(\underline{x})$ for the enlarged space is simply the same function with the $\bar{u}_j^a(x_j)$ scales now extended beyond the range 0–1 to incorporate the omitted levels. Thus, no new or additional independence assumptions are required. Finally, the method requires the same number and type of measurements and so does not introduce any new measurement error; however, it may be more sensitive to existing measurement error. That is, since $x_j^a$ and $x_j^b$ are closer together in preference than $x_j^*$ and $x_j^0$, their difference may have greater proportional error and this in turn would lead to greater error in the final $u^h(\underline{x})$ values. A measure of the extent to which this may be a problem in any particular application is the extent of extrapolation

beyond the reduced multi-attribute space required to incorporate the enlarged space. In our study, this is the value of $\bar{u}^a(\underline{x}^0)$ for the individual. For the 21 cases in our study this value ranged from 1.00 to 1.30 with a mean of 1.05. Thus, the average extrapolation was 5% indicating that, even when $x_j^b$ was not the worst level, it was a near worst level and consequently there was little impact on model error. Although the common corner points were no problem in this study, they could pose a problem in another study if the extrapolation is greater, both because of the magnifying effect on measurement error and because the independence conditions are most apt to be violated at the extremes of the ranges.

*States Worse than Death*

One of the more unusual and interesting results to come from our study is the frequency with which states are perceived as worse than death (Table IV). Eighty percent (70/87) of the subjects identified one or more health states as worse than death, and every health state was identified by some subjects as worse than death. This is the third time, to our knowledge, that social preference measurement techniques have discovered chronic health states widely considered as worse than death. Rosser and Kind (see their Table II) report similar findings for severe chronic dysfunctional states in adults, and Lathrop and Watson [1982] report similar findings for mutations. The implications of these findings of states worse than death are discussed in detail in Torrance [1982].

*Reliability*

The reliability (reproducibility) of the fundamental measurements (Table V) is consistent with our previous work in the field—correlation coefficients are relatively high, and yet so are the standard errors of measurement. The correlation coefficients range from 0.86 to 0.94 and compare favorably with previous studies where they ranged from 0.77 to 0.96 (Torrance [1976b], Torrance et al. [1973]). Precision as a proportion of scale length ranges from 0.093 to 0.153, compared to previous work where the range is from 0.081 to 0.139 (Torrance [1976b]). The high correlation coefficients suggest that, in repeated measures of health state preferences on the same individual, good states remain good and bad states remain bad. The sizeable standard errors of measurement remind us that the numerical quantification of these preferences at the individual level is not particularly precise. On the other hand, at the group level the precision is satisfactory, as demonstrated by the small standard errors of the mean (Tables III and IV).

*Model Form and Multivariate Risk Attitude*

The parameter values in Table VI show that the additive form of the

multi-attribute disutility function is inappropriate for all of our subjects. This is consistent with other work in the health field. Giauque and Peebles [1976] assert that the additive model is inappropriate for their application while Krischer [1976] found that only 10% of his 100 subjects met the conditions for additive independence.

Table VI also shows that for 91% (79/87) of the subjects, the disattributes are substitutes (attributes are complements). The intuitive notion of this is that each disattribute—P6, R5, S4 and H8—is bad and two or more together is not much worse. This finding is consistent with the assertion used by Giauque and Peebles to rule out additive independence. At first glance, the finding appears to be contrary to Krischer's result where he found speech and cosmetics to be substitutable attributes (complementary disattributes) in 87% (78/90) of the subjects with a multiplicative model. On further examination, however, both results can be seen to be examples of multivariate risk seeking behavior for losses and multivariate risk aversion for gains (Fischer and Kamlet [1981]). This occurs because of the different reference level or status quo in the two studies. In our study the subjects were healthy and were viewing the various (unhealthy) outcomes as losses; in Krischer's work the reference level was a child born with a cleft lip and palate and the various outcomes of different treatment approaches were seen as gains.

*Social Aggregation*

We investigated two methods of social aggregation—aggregation at the final level leading to $U_1(\underline{x})$ and aggregation at the basic level leading to $U_2(\underline{x})$. The former method, leading to $U_1(\underline{x})$, is considered the "correct" method for a number of reasons; it is consistent with the usual method of determining the utility of a health state for each individual and aggregating these, it is consistent with classical multi-attribute utility theory which is developed for the case of a single decision maker (or a single decision making unit), and it allows different multi-attribute utility functions (additive, multiplicative) to be used for different people as appropriate to their underlying utility structure. The other method, leading to $U_2(\underline{x})$, is considered a convenient approximation method. It is convenient because it is far less work, the results can be displayed much more compactly, and it is better suited to sensitivity analysis, to modification (like adding a new level to an attribute) and to interpretation. The question then is, how well does $U_2(\underline{x})$ approximate $U_1(\underline{x})$? In our study the approximation is relatively good; the correlation is excellent, there is a small consistent underestimate which could be added to $U_2(\underline{x})$ as a correction factor, and the remaining error appears to be nonsystematic.

*Independence Conditions*

The results from the MAU model underestimate the mean measured

utility for the two central test points F and G. There are a number of possible explanations for the underestimate; unfortunately, this study is unable to discriminate among them. The lack of fit may represent errors in the model, errors in the measurement of the test states, or both. Errors in the model, in turn, may be due to violations of mutual utility independence or to systematic errors in the measurements of the model's parameters. Violations of mutual utility independence would imply that a model more complicated than the multiplicative should be used. Alternatively, a similar result can be achieved by adding more curvature in (13). In this study the fit is improved as the exponent in (13) is increased. This suggests that perhaps the parameter value of 1.6 in (13), taken from previous work, understates the true parameter value for this study.

Systematic errors in the measurement of the model's parameters could also account for the difference between the model and the test states. A systematic error may be introduced during the measurement of the multiattribute health states A–D Table IV. These states are described to the subject by one dysfunction level and three blanks—the blanks are intended to represent "no dysfunction," but are described to the subject as being "in every other way as healthy as possible." To the extent that respondents mentally fill in the blanks with other dysfunctions, the $c_j$ values would be overstated. This would lead to a model which overstates disutility (understates utility), as observed. Also, the preference measurement procedure for states dispreferred to death is new and its validity is yet to be established. If it contributes measurement error, it would affect both the test state F and G and the MAU model.

Finally, the lack of fit may represent errors in the measurement of the test states. Fischer's review of the relevant psychological literature suggests that at the individual level discrepancies between direct holistic assessments (like states F and G) and decomposed evaluation procedures (like our MAU model) are, at least in part, due to the unreliability of the holistic assessments, especially when the number of value-relevant attributes exceeds six (Fischer, pp. 474–475). This explanation may apply in our study since, although nominally there are only four attributes, it can be argued that because three of these contain a double content, there are actually seven value-relevant characteristics which the subject must consider in making a holistic assessment. Fischer's review suggests that, in this case, the MAU model for the individual may indeed be more valid. On the other hand, the force of this explanation is mitigated because we pool results across 87 subjects and, in this way, substantially reduce the unreliability problem.

In summary, the discrepancy between the test states and the model may be due to errors in the measurement of the test states, errors in the model, or both. The current research cannot discriminate among these

possibilities, although it seems likely that it is caused by some of both. On the other hand, the current work does allow us to estimate a range of uncertainty for the utility values.

*Sensitivity Analysis*

Uncertainty about the correct utility value to assign to any health state

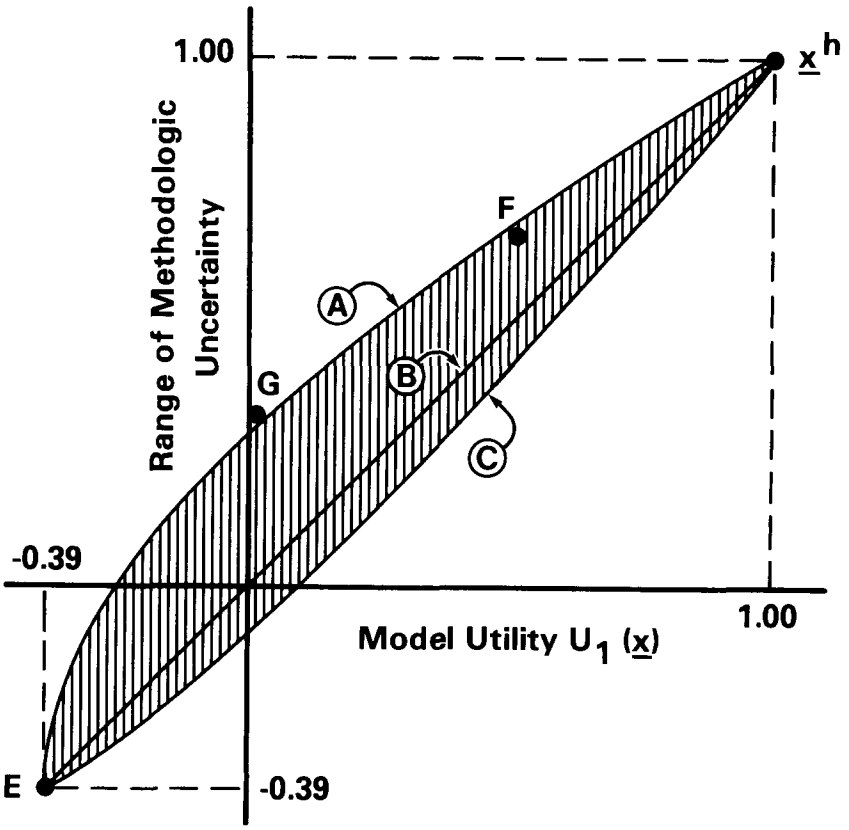


**Figure 1.** Range of methodologic uncertainty for utility values.

can be attributed to two sources, methodology and measurement. Methodology refers to our uncertainty about the true multi-attribute utility model for this situation. A range of methodologic uncertainty is shown in Figure 1. Curve B is the utility function $U_1(\underline{x})$ given in (17), which includes relationship (13). Curve C is the same utility function, but calculated without (13); i.e., assuming $\bar{u} = \bar{v}$. Finally, curve A is based on the assumption that the test states are correct and is a power curve fitted to the points. Measurement uncertainty includes both sampling error and

measurement imprecision. These are combined in the standard error $S_{\bar{x}}$ figures shown in Table IV. The mean of the seven values, 0.06, is used as a representative $S_{\bar{x}}$ value. In our study of neonatal intensive care (Boyle et al.) the two sources of uncertainty are added in order to test our findings over an extreme range of utility values. The upper limit for the sensitivity analysis is curve A + $2S_{\bar{x}}$ (not to exceed 1.0) while the lower limit is curve C − $2S_{\bar{x}}$.

## 5. CONCLUSIONS

In this study a modified MAU method was used to measure social preferences for a system of health states. The utilities so measured are for use in a cost-effectiveness (cost-utility) analysis of neonatal intensive care. Conclusions from the study are of two types: conclusions with respect to the measured utilities and their applicability in the cost-effectiveness study, and conclusions with respect to the method and its applicability by other investigators in other studies.

The measured utilities are based on responses from a random sample of parents—losses from the sample (ineligible subjects, nonparticipants and confused responders), although always undesirable, are consistent with previous studies in the field. The reliability of the basic measurements is satisfactory and comparable to previous work. The technique of prespecifying good and bad reference levels for each attribute ($x_j^a$ and $x_j^b$) and using a common set of "corner" states was successful and created only a minimal "extrapolation" of the multi-attribute utility functions. The additive form of the multi-attribute utility function is not appropriate, the multiplicative form is recommended. The two aggregation methods $U_1(\underline{x})$ and $U_2(\underline{x})$ produce comparable results—$U_1(\underline{x})$ is the theoretically correct model, and $U_2(\underline{x})$ is a convenient and close approximation. The utility correction relationship given by (13) is found to be directionally correct. $U_1(\underline{x})$ with the utility correction relationship still underestimates the two test states, and an analysis of this underestimate leads to the inclusion of the test states in the specification of a range of uncertainty on the final social preference function. The final recommended social preference function is $U_1(\underline{x})$ developed using (13) and applied in conjunction with sensitivity analysis over a specified range of uncertainty. This recommended function is used in our evaluation of neonatal intensive care (Boyle et al.).

The overall conclusion with respect to the modified MAU method is that it looks promising as an approach to the measurement of social preferences for health states that are defined by a multi-attribute classification system. Except for very small classification systems, it is a relatively efficient method compared to other approaches. It showed good potential in this study and no insurmountable difficulties were encoun-

tered. However, a number of lessons were learned, and these are summarized below as recommendations to other investigators who may be interested in reapplying the method in other studies. First, the health state classification system should be designed such that each attribute contains only one concept (to minimize respondent confusion), and such that the attributes can be conceived of by the subjects as being independent. It is not essential that the attributes actually *be* independent in reality all the time, but simply that the subjects are able to *visualize* them as independent. This is required because each measurement question must specify what the subject is to assume about the other attributes, and this assumption must be plausible. The preference measurement instruments were acceptable to the subjects and showed satisfactory reliability. They can be recommended for reuse in other studies with two provisions: first, it would be advisable to recalibrate the $\bar{u} = \bar{v}^{1.6}$ relationship for the particular situation under study and second, for states dispreferred to death, it would be useful to measure the utility by several techniques to investigate validity. The technique of prespecifying common "corner" states was successful in this study and can be recommended for reuse. With respect to the use of test states to check the fit of the model our recommendation is twofold. For problems with many attributes (about seven or more) the use of test states is infeasible because of the unreliability of the holistic utility assessments of the test states themselves. In such cases the necessary independence conditions must be investigated directly. For problems with few attributes (about six or less) the use of test states is a feasible option, but they should be randomly selected from the entire multi-attribute space and there should be a sufficient number to be properly representative. Although our study found the multiplicative model to be more appropriate than the additive model for all subjects, this cannot be generalized. Each study will have to determine from its own data which model is more appropriate for each subject and overall. In our study the two methods of aggregation produced comparable results and, although we suspect that this will apply to other studies, we recommend, for the minimal extra work involved, that other researchers investigate this in their own situation. Finally, we hope that other investigators interested in measuring social preferences for health states which can be defined by a multi-attribute classification system will consider and use the MAU method along with the modifications and recommendations they find appropriate.

## NOTES

1. Relationship (13) was not reported directly in that form in our previous work. The actual reported functions in the original notation ($C$ = category scaling, $T$ = time trade-off, $S$ = standard gamble) are $C = 1 - (1 - T)^{0.62}$ (Torrance

[1976b]) and $T = S$ (Torrance [1976b], Torrance et al. [1973]) which combine to form $(1 - S) = (1 - C)^{1.61}$. In addition, our previous work (Torrance [1976b]) found but did not report the relationship $(1 - T) = (1 - C)^{1.58}$ which would lead to the same relationship (13).

2. Table II in Wolfson et al. [1982] provides 35 data points relating $S$ to $C$. Although Wolfson fits a linear function to this data, a power curve like (13) can also be fitted giving $(1 - S) = (1 - C)^{2.16}$. Thus, the Wolfson data confirm the same general shape of the relationship although certainly not the specific parameter value.

## ACKNOWLEDGMENTS

This research was supported in part by the Ontario Ministry of Health (DM366) and by the Natural Sciences and Engineering Research Council (A4129). The authors also thank John Sinclair, Kathy Bennett, Jim Julian and an anonymous referee for their help.

## REFERENCES

BOYLE, M. H., G. W. TORRANCE, J. C. SINCLAIR AND S. P. HORWOOD. 1982. Economic Evaluation of Neonatal Intensive Care of Very Low Birth-Weight Infants, McMaster University, Hamilton, Ontario.

DYER, J. S., AND R. K. SARIN. 1979. Group Preference Aggregation Rules Based on Strength of Preference. *Mgmt. Sci.* **25,** 822–832.

FANSHEL, S., AND J. W. BUSH. 1970. A Health Status Index and Its Application to Health Services Outcomes. *Opns. Res.* **18,** 1021–1066.

FARQUHAR, P. H. 1977. A Survey of Multiattribute Utility Theory and Applications. In *Studies in Management Science; Vol. 6. Multiple Criteria Decision Making*, pp. 59–89, Martin K. Starr and Milan Zeleny (eds.). North-Holland, Amsterdam.

FISCHER, G. W. 1979. Utility Models for Multiple Objective Decisions; Do They Accurately Represent Human Preferences? *Decision Sci.* **10,** 451–479.

FISCHER, G. W., AND M. S. KAMLET. 1981. The Reference Level Risk-Value Model: An Expected Utility Analysis of Reference Effects and Multivariate Risk Preferences. Paper presented at CORS/ORSA/TIMS Joint National Meeting, Toronto, May 3–6. (Address correspondence to the Department of Social Science, Carnegie-Mellon University, Pittsburgh, PA 15213.)

GIAUQUE, W. C., AND T. C. PEEBLES. 1976. Application of Multidimensional Utility Theory in Determining Optimal Test-Treatment Strategies for Streptococcal Sore Throat and Rheumatic Fever. *Opns. Res.* **24,** 933–950.

HARSANYI, J. 1975. Nonlinear Social Welfare Functions. *Theory and Decision* **6,** 311–332.

HILDRETH, C. 1953. Alternative Conditions for Social Orderings. *Econometrica* **21,** 81–94.

KALAI, E., AND D. SCHMEIDLER. 1977. Aggregation Procedure for Cardinal Preferences: A Formulation and Proof of Samuelson's Impossibility Conjecture. *Econometrica* **45,** 1431–1438.

KAPLAN, R. M., J. W. BUSH AND C. C. BERRY. 1976. Health Status: Types of Validity and the Index of Well-being. *Health Serv. Res.* **11,** 478–507.

KAPLAN, R. M., J. W. BUSH AND C. C. BERRY. 1979. Health Status Index: Category Rating versus Magnitude Estimation for Measuring Levels of Well-being. *Med. Care* **17,** 501–525.

KEENEY, R. L. 1976. A Group Preference Axiomatization with Cardinal Utility. *Mgmt. Sci.* **23,** 140–145.

KEENEY, R. L., AND H. RAIFFA. 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs.* Wiley, New York.

KRISCHER, J. P. 1976. Utility Structure of a Medical Decision-Making Problem. *Opns. Res.* **24,** 951–972.

LATHROP, J. W., AND S. R. WATSON. 1982. Decision Analysis for the Evaluation of Risk in Nuclear Waste Management. *J. Opnl. Res. Soc.* **33,** 407–418.

VON NEUMANN, J., AND O. MORGENSTERN. 1953. *Theory of Games and Economic Behavior.* Wiley, New York.

PATRICK, D. L., J. W. BUSH AND M. M. CHEN. 1973a. Toward an Operational Definition of Health. *J. Health Soc. Behav.* **14,** 6–23.

PATRICK, D. L., J. W. BUSH AND M. M. CHEN. 1973b. Methods for Measuring Levels of Well-being for a Health Status Index. *Health Serv. Res.* **8,** 228–245.

RICHARD, S. F. 1975. Multivariate Risk Aversion, Utility Independence, and Separable Utility Functions. *Mgmt. Sci.* **22,** 12–21.

ROSSER, R., AND P. KIND. 1978. A Scale of Valuations of States of Illness: Is There a Social Consensus? *Int. J. Epidemiol.* **7,** 347–358.

ROSSER, R., AND V. WATTS. 1978. The Measurement of Illness. *J. Opnl. Res. Soc.* **29,** 529–540.

SACKETT, D. L., AND G. W. TORRANCE. 1978. The Utility of Different Health States as Perceived by the General Public. *J. Chronic Dis.* **31,** 697–704.

TORRANCE, G. W. 1976a. Health Status Index Models: A Unified Mathematical View. *Mgmt. Sci.* **22,** 990–1001.

TORRANCE, G. W. 1976b. Social Preferences for Health States: An Empirical Evaluation of Three Measurement Techniques. *Socioecon. Planning Sci.* **10,** 129–136.

TORRANCE, G. W. 1982. Multi-Attribute Utility Theory as a Method of Measuring Social Preferences for Health States in Long-Term Care. In *Values and Long-Term Care,* pp. 127–156, R. Kane and R. Kane (eds.). Lexington Books Div., D. C. Heath Publisher, Lexington, Mass.

TORRANCE, G. W., W. H. THOMAS AND D. L. SACKETT. 1972. A Utility Maximization Model for Evaluation of Health Care Programs. *Health Serv. Res.* **7,** 118–133.

TORRANCE, G. W., D. L. SACKETT AND W. H. THOMAS. 1973. Utility Maximization Model for Program Evaluation: A Demonstration Application. In *Health Status Indexes,* pp. 156–165, R. L. Berg (ed.). Hospital Research and Educational Trust, Chicago.

WEINSTEIN, M. C., AND W. B. STASON. 1977. Foundations of Cost-Effectiveness Analysis for Health and Medical Practices. *N. Engl. J. Med.* **296,** 716–721.

WOLFSON, A. D., A. J. SINCLAIR, C. BOMBARDIER AND A. MCGEER. 1982. Preference Measurements for Functional Status in Stroke Patients: Inter-Rater and Inter-Technique Comparisons. In *Values and Long-Term Care,* pp. 191–214, R. Kane and R. Kane (eds.). Lexington Book Div., D. C. Heath Publisher, Lexington, Mass.