METHODOLOGY FOR MEASURING HEALTH-STATE PREFERENCES—I: MEASUREMENT STRATEGIES

DEBRA G. FROBERG* and ROBERT L. KANE

Division of Human Development and Nutrition, School of Public Health,
University of Minnesota, Minneapolis, MN 55455, U.S.A.

(Received in revised form 25 July 1988)

Abstract—Values play a critical part in decision making at both the individual and policy levels. Numerous methodologies for determining the preferences of individuals and groups have been proposed, but agreement has not been reached regarding their scientific adequacy and feasibility. This is the first of a four-part series of papers that analyzes and critiques the state-of-the-art in measuring preferences, particularly the measurement of health-state preferences. In this first paper we discuss the selection of relevant attributes to comprise the health-state descriptions, and the relative merits of three measurement strategies: holistic, explicitly decomposed, and statistically inferred decomposed. The functional measurement approach, a statistically inferred decomposed strategy, is recommended because it simultaneously validates the process by which judges combine attributes, the scale values they assign to health states, and the interval property of the scale.

INTRODUCTION

We all need to make decisions about health care. Regardless of the position one occupies within the health care system—as patient, consumer, health provider, or policymaker—the complexity of information and the difficulty of making choices can be overwhelming. A patient may have to decide whether to undergo a painful treatment that has a high probability of prolonging life, but will considerably decrease the quality of his or her remaining years. Even those of us who are well face choices such as what type of health insurance to purchase. Both employers and individuals must evaluate the relative merits of HMOs, PPOs, and commercial health insurance plans, plans that may differ in cost, freedom of choice, amenities, and even quality of care. As health care costs continue to escalate, policymakers also confront tough decisions about what programs to fund and how widely available to make them. Since not all worthy programs can be funded, how should one decide between, say, community services for the elderly and prenatal care for low income women? Although the nature of these decisions is quite different, choices among treatments, health plans, and policies all involve evaluating options on the basis of their likelihood of bringing about outcomes we value. Thus, a critical element of decision making is determining what we value.

While determining values may seem at first blush a reasonably easy thing to do, further reflection reveals a web of difficulties. Choices are rarely black and white. More often than not, they involve trading one desirable (or undesirable) outcome for another, as when a patient accepts the side effects of a drug in order to reduce his or her risk of a stroke. Moreover, values are not static but may change over time or in response to specific experiences. Even more complex than incorporating values into clinical decisions for a single patient is the use of values in setting policy. Collective decision making involves additional issues such as how to elicit values from appropriate constituencies and how to aggregate values in a way that is both techni-
cally defensible and morally just. Thus, it is not surprising that while the importance of measuring individual preferences is well recognized among some professionals, especially those familiar with methods of decision analysis, using patients' and society's values in decision making is far from common practice [1].

In this paper we focus on the measurement of individual preferences, deferring some of the philosophical questions associated with using preferences until we have determined whether accurate, reliable, and feasible methods for measuring values exist. We further restrict our domain to the measurement of preferences for health states since that has been the focus of the bulk of the work in preference measurement. Extrapolation to the other areas may be possible, but it must be undertaken with caution. At this juncture we will drop the word "values" and use only the terms "preferences" or "utilities". Although all three words are often used interchangeably in the literature, clarity will be enhanced by defining values as the more general dispositions which serve as a basis for preferences. In this paper, preferences or utilities refer to levels of subjective satisfaction, distress, or desirability that people associate with a particular health state. Other synonyms for this level of subjective satisfaction are quality of life, weight, or rating of the health state [2].

In general, various approaches to obtaining health-state preferences have included these three steps: (1) defining a set of health states of interest, (2) identifying a judge or group of judges to provide judgments of the desirability of each health state, and if necessary, (3) aggregating across the judges to determine scale values for each health state [3].

Within this general framework, however, the researcher must make a series of decisions about how to proceed. These decisions have been discussed in the literature but controversy still surrounds each one. In this series of papers, we scrutinize the literature relating to the measurement of health states. We pay particular attention to the following unresolved questions:

1. What are the relevant health dimensions?
2. How should health states be presented to the respondents? (For example, should respondents rate each health dimension separately, or should they rate holistic health states composed of multiple dimensions?)
3. What preference scaling method (e.g. standard gamble, rating scale) should be used?
4. Do population groups (e.g. general public, health care professionals, patients) differ in their preferences?
5. How can situational variables be controlled in order to make preference values more consistent and accurate?

These four papers are based on a comprehensive search of literature published over the past 20 years. The search strategy began with a MEDLINE search consisting of five steps: (1) specifying that the major focus of the article should be "health status indicators" and that it also had to be about "methods"; (2) pairing "health status indicators" with each of the following: social perception, self-concept, decision theory, decision making, choice behavior, and judgment; (3) identifying articles in which "health status" or "preferences" appeared in the title or abstract; (4) using the medical headings "health status", "health status indicators", and "attitudes to health" to select articles in which the words "preference" or "perception" appeared in the title or abstract; (5) pairing "attitude to health" with each of the following: perception, methods, values, and preferences; and (6) selecting four well-published authors and pairing their names with "health status" and "health status indicators".

Beyond the formal computer search, we obtained additional books and articles by consulting the reference lists of articles generated by the search, and by perusing journals most likely to contain relevant articles. Personal communication with investigators working the field yielded several more articles and some unpublished material.

SELECTING RELEVANT HEALTH DIMENSIONS

When developing health-state descriptions, the purpose of the research dictates the types of attributes to be included. For some purposes, a comprehensive set of attributes is required, whereas for other purposes, a more restricted set of attributes will suffice. A rule of thumb is that no more than nine attributes and preferably fewer should be used since research consistently has shown that humans can process simultaneously only five to nine pieces of information [4]. Attributes are most commonly chosen on the basis of conceptual considerations, but some investigators have incorporated consensus data from clinical experts [5] or data obtained from patients or patients' relatives [6]. Examples of health attributes are physical function, social
### Table 1. Example of a health-state classification system

<table>
<thead>
<tr>
<th>Mobility</th>
<th>Pain</th>
<th>Emotional well-being</th>
</tr>
</thead>
<tbody>
<tr>
<td>No limitations</td>
<td>No pain</td>
<td>Not depressed</td>
</tr>
<tr>
<td>Walks with a limp</td>
<td>Mild pain</td>
<td>Slightly depressed</td>
</tr>
<tr>
<td>Uses a crutch or aid</td>
<td>Moderate pain</td>
<td>Moderately depressed</td>
</tr>
<tr>
<td>Does not walk</td>
<td>Severe pain</td>
<td>Very depressed</td>
</tr>
</tbody>
</table>

Adapted from Boyle and Torrance [7].

function, emotional well-being, pain, and cognitive ability [7]. For each attribute, a number of levels can be defined which represent stepwise increments from good to poor functioning. The description of each level generally focuses on function rather than on clinical diagnosis. A simple example is shown in Table 1.

Health states are usually formed by taking one level from each attribute. In this example there are four mobility levels, four pain levels, and four emotional well-being levels. Thus, there are $4 \times 4 \times 4 = 64$ potential combinations of levels, or 64 potential health states. Each different health state has a potential value associated with it, alternatively referred to in the literature as a weight, a cardinal value (the word "cardinal" refers to a value that has equal-interval or ratio properties) or index value. Obtaining those values is the central problem addressed in this paper. To derive cardinal values for each unique health state, the investigator must first decide how to present health states to respondents for evaluation. We will call this the measurement strategy.

### SELECTING A MEASUREMENT STRATEGY

There is some confusion in the literature over the distinction between measurement strategy and scaling method. This confusion is heightened by differences in terminology used by various investigators in referring to the same concepts. In this paper, measurement strategy refers to the overall structure for posing questions to the respondents (e.g. having respondents rate multiattribute health states vs rating each attribute separately) and the corresponding method of analyzing the data (e.g. regression analysis, analysis of variance). On the other hand, the scaling method is the specific task required of the respondent to achieve scale values for health states. Many different scaling methods have been used in studies of health preferences, including the standard gamble, time trade-off, rating scale, magnitude estimation, equivalence and willingness-to-pay.

Measurement strategy considerations must logically precede scaling method considerations. Besides determining the kinds of questions that will be posed to the respondent, the measurement strategy also specifies the kinds of hypotheses that can be tested and thus the kinds of conclusions that can be drawn from the data [8]. The investigator’s choice of measurement strategy has a major impact on the amount of information that can be given to rating judges, particularly on the number of value judgments required from each judge. Choice of a scaling method will also depend upon constraints imposed by the measurement strategy [9].

Fischer [10] presents a framework for classifying measurement strategies from the perspective of multiattribute utility theory. Veit and colleagues [11] discuss similar concepts from a psychometric perspective. Both of these excellent reviews are drawn upon here. Broadly speaking, two general approaches have been applied to measuring preferences for health states. The holistic approach requires the judge to assign scale values to each possible health state, when a health state represents a combination of many attributes. This may be accomplished using any of the aforementioned scaling methods, (rating scales, standard gamble, etc.). On the other hand, the decomposed approach enables the investigator to obtain values for all health states without requiring the judge to assign values to every one. It simplifies the assessment task by expressing the overall value of a health state as a decomposed function of the attributes. This will be discussed in detail later, but for now it is important to note that decomposed scaling methods can greatly reduce the number of subjective judgments required to assign scale values to a complete set of health states [10].

**Holistic Designs**

All of the early pioneering work in the measurement of preferences for health states has used the holistic approach [3]. This strategy requires respondents to rate each multiattribute health state of interest to the investigator; however, separate effects of each attribute are not analyzed.

Two examples illustrate variations in the way this strategy has been applied. Patrick and colleagues [12] defined 29 function levels, five age groups, and 42 symptom/problem complexes. The function levels were a composite of three attributes: physical activity, mobility, and
social activity. Next, they combined the function levels, age groups and symptom/problem complexes to form a matrix for describing the universe of conditions that may affect the health status of a population. From this complete set of combinations (which numbered in the thousands), 400 case descriptions were sampled and given to judges to evaluate using the rating scale method. (Not all 400 case descriptions were given to all judges.) A minimum of 10 items were chosen at each function level by using a random number table to sample symptom complexes at each level and to sample age groups within each complex. For example, a case description read as follows:

- 6–17 years.
- Walked freely.
- Travelled freely.
- Did not perform major activity but performed self-care activities.
- Had cough, wheezing, or shortness of breath.

Patrick then computed an average rating of the sampled items at each function level and these became scale values for each level.

A second early study using the holistic approach [13] developed more detailed health scenarios than those used by Patrick and colleagues. Sackett and Torrance [13] chose 10 well-understood disorders such as depression, hospital dialysis, and mastectomy for breast cancer, and developed scenarios describing the physical, social, and emotional characteristics of each state. Scale values for each state were determined through a time trade-off technique, which presents the judge with two scenarios than those used by Patrick and others. The scenario for dialysis was as follows:

"You often feel tired and sluggish. A piece of tubing has been inserted into a vein in either your arm or your leg. This might restrict some of your physical movements. There is no severe pain, but rather chronic discomfort. Two or three times each week you must go to the hospital and spend about 6-17 hours hooked up to a dialysis machine. If your job does not involve strenuous physical labour, you may continue to work and undergo dialysis at night. You must follow a strict diet: low salt, little meat, and small amounts of fluid. You are free to travel about your community but further travel is restricted by the necessity to return to your dialysis machine.

Many people become depressed with the nuisances and restrictions which have become part of their lives. Also, there is the knowledge that you are being kept alive by the machine.

Any limitations on your social life would be due to your feeling tired, dietary restrictions (very little drinking), and the time that must be spent in the hospital. Your activities must be scheduled around your visits for dialysis." [13, p. 698]

Investigators using the holistic approach often assume that the scale values have equal interval properties.* A major limitation of holistic approaches is that the assumption of equal intervals is based on definition rather than on an empirically verified hypothesis, and the holistic strategy makes it impossible to test the hypothesis. Additional studies, seldom conducted in practice, are necessary to test the validity of the assumption. A second limitation of holistic designs is that they do not provide information about how the different attributes are weighted and combined to produce the values associated with each multiattribute health state [8]. Further, the burden placed on judges to rate a large number of multiattribute health states restricts the applicability of these approaches. For these reasons, holistic strategies are being replaced by decomposed methods in more recent studies of health-state preferences.

Decomposed Designs

In contrast to holistic designs, decomposed designs greatly reduce the number of subjective judgments required to assign scale values to a complete set of health states. Within the general category of decomposed designs, one can distinguish between (1) assessment procedures that attempt to develop an algebraic model of the decision maker's preferences from a set of multiattribute judgments (statistically inferred models), and (2) assessment procedures that permit the decision maker to break up the overall evaluation process into a set of simpler subtasks (explicitly decomposed models) [10]. Like the holistic approach discussed above, the algebraic modeling approach requires the respondent to rate multiattribute health states. However, algebraic modeling differs from the holistic approach in that it does not require that all multiattribute health states be evaluated. It also allows the attributes comprising the health states to be separated and their individual effects analyzed. This feature is important in that it provides information about how judges combine the attributes to arrive at an overall judgment.

*In an interval scale, any two categories have magnitudes that are separated by a measurably equal interval. This property is important for the application of some statistical analyses, and in cost-effectiveness analysis.
Explicitly decomposed models

Explicit decomposition procedures ask the respondent to evaluate each level of a particular attribute assuming all other attributes are held constant. Thus, they require few (and in some cases no) multiattribute judgments. While there are numerous variations within this approach, only one, the conditional utility function-based procedure, will be discussed here.

The general class of explicitly decomposed models constitutes the standard multiattribute utility (MAU) method. MAU theory originated in the early 1960s as decision analysts from a variety of disciplines recognized the need to expand methods of decision analysis to situations in which the decision maker is faced with multiple, competing objectives rather than a single, well-defined objective. MAU theory is concerned with the construction of multiattribute utility functions. It specifies several possible functions (additive, quasi-additive, and multilinear) and the independence conditions under which each would be appropriate [3]. The establishment of these conditions makes it possible to represent utilities for multiattribute states using explicit decompositional procedures. This means that rather than having to rate multiattribute health states, the judge can rate each attribute separately. The conditional utility function method involves three major subtasks:

1. Checking independence assumptions to determine which—if any—of the decomposed model forms is appropriate,
2. Assessing utility functions over single-outcome attributes, and
3. Measuring the utility of selected multiattribute health states to determine scaling constants, thereby permitting aggregation of utility over attributes [10].

The first step, checking independence assumptions, refers to independence among the attributes. That is, is the effect of one attribute (e.g., physical health) independent of the effect of other attributes (e.g., mental health)? If physical health is independent of mental health, then preferences for various states of physical health, holding mental health fixed, do not depend on the particular level at which mental health is fixed. This situation, in which there are no interactions among the attributes, is known as the additive model.

Technically, three conditions must be satisfied in order to assume an additive model: utility independence, mutual utility independence, and additive utility independence. If only the first condition is satisfied, the model is multilinear; and if the first two conditions are satisfied, the model is quasi-additive. The three conditions form a hierarchy, defined as follows:

1. Utility independence requires that each attribute is utility independent of all other attributes. This means that preferences for various levels of each attribute do not depend upon the particular levels at which the other attributes are fixed. A model satisfying only this condition is multilinear.

2. Mutual utility independence requires that every subset of attributes is utility independent of its complement (the set of remaining attributes). This means that preferences for the various levels of each subset of attributes do not depend upon the particular levels at which the remaining attributes are fixed. A model satisfying this condition in addition to condition 1 (above) is quasi-additive.

3. Additive utility independence requires that if we let the multiattribute state with all attributes at their most preferred level equal 1.0, and the multiattribute state with all attributes at their least preferred level equal 0.0; then, if each attribute takes on its most preferred value and at the same time, all remaining attributes take on their least preferred values, the sum of these utilities across attributes should equal 1.0. This means that the whole is equal to the sum of its parts, and that the contribution of each attribute is independent of the values of the remaining attributes. If this condition is satisfied in addition to the first two, the model is additive [10].

Keeney and Raiffa [14] have shown that additive utility independence implies mutual utility independence, but that the converse is not true.

There are a variety of methods for checking independence assumptions [14]. Unfortunately, because they all assume that the decision maker's utility assessments are free from random response error, the investigator must decide how large a deviation from linearity to tolerate before rejecting the utility independence assumption. (Anderson's functional measurement approach, to be discussed later, deals with this problem through the use of analysis of variance.) A second difficulty associated with the establishment and verification of independence conditions is the fact that it is "a tedious, exacting, and time-consuming task requiring..."
extensive interviewer-subject interaction”, feasible only in studies with a small number of subjects ([3] p. 1051). Thus, in practice, investigators often modify this step as did Torrance et al. [3], who elected to assume the existence of mutual utility independence and test the assumption later using judges’ holistic assessments of multiattribute health states.

After determining which of the three models (additive, quasi-additive, or multilinear) is appropriate, the investigator asks the judge to evaluate each level of a particular attribute assuming all other attributes are held constant. Usually the least and most preferred levels of any attribute are assigned the values of 0–1, and the intermediate values can be determined through the use of a scaling technique such as category rating or the standard gamble.

In the third and final step, the judge provides scaling constants by assessing utilities of selected multiattribute health states. These scaling constants can be thought of as “importance weights” for each attribute. Taken together, these three steps represent the multiattribute utility approach, and provide a means of expressing utilities of multiattribute health states as a function of the utilities of each attribute taken singly.

A good example of the explicitly decomposed multiattribute utility method using the conditional utility function-based procedure can be found in a study conducted by Torrance et al. [3]. These investigators measured preferences for health states for use in a cost-effectiveness analysis of neonatal intensive care. Several modifications in the standard multiattribute utility (MAU) method were made relative to the establishment and verification of independence conditions, scaling techniques, definition of extreme levels, and aggregation of individual preferences into social preferences. The judges, who were parents of school children, provided individual single-attribute value functions using the category scaling method. They also provided individual utilities for multiattribute states using the time trade-off technique. In their discussion of the method and the results of their study, Torrance and his colleagues conclude that the modified MAU method is a relatively efficient way of measuring health states that are defined by a multiattribute classification system. Compared to holistic designs, the MAU approach “does not provide any way to validate the weights, the utilities, or the model and thus any prescribed outcomes; nor is there any way beyond definition of knowing what the scale properties of the numbers are” (p. 253).

Statistically inferred decomposed models

Both explicitly decomposed models and statistically inferred models require the judge to make fewer subjective judgments than do holistic models. In addition, one statistically inferred technique (the functional measurement method) has the additional advantage of permitting a test of the underlying subjective processes by which respondents process information, thereby providing a validation of the derived scale values.

Functional measurement. At the heart of the functional measurement approach is the principle of simultaneously testing theories of information processing and measuring scale values. According to Anderson [15], the investigator associated with this approach, the two go hand-in-hand; subjective constructs can only be measured in the context of a valid theory.

Figure 1 illustrates a theory of human information processing. If we think of the observed stimulus information \((i \text{ and } j)\) on the left as particular levels of two attributes of a health state (say, mental and physical health), the model operates as follows: First, the respondent transforms each piece of information (e.g. severe depression, no physical limitations) contained in a health state into a subjective stimulus value \((S_i, S_j)\) by the function \(H\). The respondent then uses a combination rule \((C)\) to transform these scale values into a subjective response, \(\psi\). Finally, the respondent transforms this subjective response into an observed response, \(R\), using the function \(J\).

Measurement thus involves three simultaneous problems: (a) measuring the subjective stimulus values on equal-interval scales, (b) measuring the subjective response value on an equal-interval scale, and (c) finding the psychological law that relates the subjective values of stimuli and response. In the functional measurement approach, all three problems are solved together [15].

Solving these three problems simultaneously requires the use of a factorial design. Such a design permits a test of (c) above, the law that relates the subjective values of stimuli and response. (This corresponds to the combination
Methodology for Measuring Health-state Preferences—I

<table>
<thead>
<tr>
<th>Stimulus Information</th>
<th>Subjective Scale Values</th>
<th>Combination Rule</th>
<th>Subjective Scale Values</th>
<th>Overt Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Observed)</td>
<td>(Subjective)</td>
<td>(Subjective)</td>
<td>(Subjective)</td>
<td>(Observed)</td>
</tr>
</tbody>
</table>

Fig. 1. Outline of subjective processes. (Veit et al. [11])

rule (C) in Fig. 1.) If the data support the predictions of the model, subjective stimulus ($S_i$ and $S_j$ in Fig. 1) and response ($ρ_{ij}$) scale values can be derived from the model [8]. Suppose we have two factors (mental and physical health) and the first factor has four levels and the second factor has five levels. In a factorial design, all levels of mental health are combined with all factors of physical health to produce $4 \times 5 = 20$ possible health states.

The data produced by factorial designs are analyzed using analysis-of-variance procedures. If the data generated by respondents' evaluations of each multiattribute health state obey the conditions of the model, the model is accepted as an appropriate description of the combination process, and the stimulus and response scales are separately derived from the model. The additive model is supported if no interactions are present. If statistically significant interactions are present, procedures are available for determining whether these interactions can be described by the quasi-additive or multilinear models described earlier. If so, it is again possible to derive stimulus and response values [10]. Table 2 displays hypothetical data generated from a factorial design with a mental health factor and a physical health factor. Cell entries (the values in the body of the table) are means calculated from two-factor health-state ratings made by a group of judges. The marginal means (the values outside the table) represent the main effects of Factors A and B, mental and physical health, respectively.

Proponents of the functional measurement approach claim that it provides an extremely powerful device for validating the combination rule while at the same time validating the scale values. Unlike other methods, functional measurement methods permit conclusions about the level of measurement (i.e., ordinal, interval, ratio) of scaled health states. (This corresponds to transformation $J$ in Fig. 1.) Suppose, for example, that a set of scale values resulting from category ratings satisfies the additive model; that is, an analysis of variance shows that there are no significant interactions among attributes. That is to say, the attributes are independent so that when responses to one attribute are plotted as a function of each of the levels of the other factor, the curves are parallel as in Fig. 2. (Figure 2 is a graphic representation of the data in Table 2.)

Given this parallelism, Anderson [16] explains how the absence of interaction among attributes validates an interval level scale:

A priori, there is no great reason to think that ordinary ratings constitute an interval scale of response. However, if the overt response were a nonlinear function of the underlying response, then the data would not plot as parallel lines even if the model were true. Parallelism thus provides a joint validation of the psychological law, and of the response scale (p. 221).

Table 2. Hypothetical data from a factorial design

<table>
<thead>
<tr>
<th>Mental Health (Factor A)</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Health Level 1</td>
<td>1.0</td>
<td>3.0</td>
<td>4.0</td>
<td>5.0</td>
<td>3.25</td>
</tr>
<tr>
<td>Physical Health Level 2</td>
<td>3.2</td>
<td>5.2</td>
<td>6.2</td>
<td>7.2</td>
<td>5.45</td>
</tr>
<tr>
<td>Physical Health Level 3</td>
<td>4.1</td>
<td>6.1</td>
<td>7.1</td>
<td>8.1</td>
<td>6.35</td>
</tr>
<tr>
<td>Physical Health Level 4</td>
<td>4.6</td>
<td>6.6</td>
<td>7.6</td>
<td>8.6</td>
<td>6.85</td>
</tr>
<tr>
<td>Physical Health Level 5</td>
<td>5.0</td>
<td>7.0</td>
<td>8.0</td>
<td>9.0</td>
<td>7.25</td>
</tr>
<tr>
<td>Physical Health</td>
<td>3.58</td>
<td>5.58</td>
<td>6.58</td>
<td>7.58</td>
<td></td>
</tr>
</tbody>
</table>

Adapted from Veit and Ware [8].
The shortcoming of the functional measurement approach may occur if the number of attributes is large. Fischer [10] reviewed several studies comparing functional measurement with explicit decomposition procedures and found that with six or fewer attributes, the two methods assigned very similar values to outcomes. However, this convergence declined with larger numbers of attributes and the evidence suggested that this was due to a deterioration in the reliability of multiattribute judgments. Other investigators have found that when only a few attributes are involved, multiattribute judgments are more reliable than decomposed judgments [17, 18].

Although the functional measurement approach is new to health services research, it has been applied in several studies of health-state preferences. Veit and colleagues [11] constructed 16 different health states by combining two attributes: four levels of a physical attribute and four levels of a mental attribute. They found that there were systematic interactions between physical and mental health attributes, so that when health was poor on one component, the other component had less effect. In contrast, Cadman and Goldsmith [9] found no significant interactions among the eight attributes they examined. Their study used a factorial design to develop a function index for evaluating a program for the care of young handicapped children. It is a good example of how fractional factorial designs can be used to reduce respondent burden when the number of attributes and levels is large. In a third study, patients’ values for three aspects of voice function were assessed prior to and following radiotherapy for laryngeal cancer. Like Cadman and Goldsmith, the investigators found that a simple additive model with no interactions provided a good fit to the data [17].
additive model) using regression procedures [10]. Conclusions concerning the adequacy of the model are based on the magnitude of the multiple correlation coefficient. If $R^2$ is about 0.7 or 0.8 it is usually concluded that the degree of correspondence between the model-generated utilities and the judges' multiattribute evaluations is acceptably high. Regression analysis rests upon two assumptions: that the stimulus values are known, and that the overt response is on an equal-interval scale.

The main problem with this approach is that it does not test the validity of the scale values. Because investigators generally employ direct scaling procedures to obtain scale values, the validity of these input values is unknown. Whereas the functional measurement approach incorporates scaling as an integral part of testing the underlying information-processing theory and thus validate scale values along with the theory, regression techniques do not provide a way to determine the validity of the scale values. Regression techniques simply assume they are valid and use these input values to test the combination rule. The multiple correlation coefficient ($R^2$) is not an adequate test of scale values because $R^2$ can be high even when deviations from model predictions are significant and systematic [8]. Reported regression analyses seldom include a test of the fit of the linear regression model, even though it has been demonstrated that important interactions can exist even with an $R^2$ as high as 0.98 for an additive model [15]. Finally, the fact that stimuli are often intercorrelated further obscures the meaning of the multiple correlation coefficient. For more indepth discussion of the use of multiple regression procedures for determining subjective values, the reader is referred to Wiggins and Hoffman [19], Anderson [15], Huber et al. [20], Hoepfl and Huber [21], and Birnbaum [22, 23].

**SUMMARY AND CRITIQUE**

The first two issues the investigator of health-state preferences must address are selecting relevant health attributes and selecting a measurement strategy. Selection of attributes will depend upon the investigator's purpose, but a general rule is to use nine or fewer attributes if multiattribute judgments are to be made. We have used the term "measurement strategy" to describe the structure that determines how questions will be posed to the respondent, and what kinds of conclusion can be drawn from the data.

Two broad classes of strategies were discussed. Holistic strategies have been used extensively in the past but are gradually being replaced by decomposed strategies. Decomposed strategies may be classified as either explicitly decomposed or statistically inferred. The principal virtue of decomposed strategies is that they require fewer subjective judgments, a particular advantage when the number of attributes is large. In addition, one type of statistical approach, functional measurement, permits a test of the underlying information-processing theory. From a technical standpoint, the functional measurement approach is clearly superior to the other designs discussed in this paper. It is the only approach that simultaneously validates the process by which judges combine attributes, the scale values they assign to health states, and the interval property of the scale. Although a few studies have successfully used the approach, the practicality of functional measurement remains to be seen. Finally, it is prudent to limit the number of attributes to nine or fewer, since judgments of multiattribute health states containing more than nine attributes are likely to be invalid.

The measurement strategies discussed in this paper implicitly assume that individual preferences can be aggregated to form social preferences by simply calculating the arithmetic mean. It should be emphasized that while this paper is devoted to measurement issues, anyone contemplating aggregating individual health-state preferences for purposes of program evaluation or policy analysis should be aware of the literature in the area of social choice theory. Whether and how to aggregate individual preferences have been the subject of much debate among welfare economists and philosophers ever since the publication of Arrow's Impossibility Theorem. Arrow [24] showed that no social welfare function, i.e. method of developing a group choice as an aggregation of preferences of its members, can satisfy four reasonable assumptions. Later it was shown that when cardinal utilities are used instead of rankings, it is possible to define consistent aggregation rules; however, these rules explicitly require interpersonal comparison of preference [25].

The appropriateness of making interpersonal comparisons of utility lies at the heart of the controversy over aggregating preferences. Resnick [26], for example, describes how individuals' scales can be recalibrated such that a unit on one person's scale is the same as a unit
on another person's scale. On the other hand, Torrance [3] handles the problem by establishing two clearly defined outcomes, one good and one bad, as anchor points, but not necessarily end points, for the utility scale. The central basis for aggregation is that the difference in utility between these two outcomes of "a normal healthy life" and "death" is set equal across people. In addition to the controversy surrounding interpersonal comparison of utility, using the arithmetic mean raises questions of equity, since the same mean value can arise if, for example, three people all give a health state a rating of 20 utility points as when two people give it 30 utility points and one person gives it 0 points. These issues cannot be thoroughly discussed and resolved here, but they should be considered whenever preferences are aggregated for applied purposes.

Acknowledgements—The authors wish to express their appreciation to Allan Detsky, Walter Spitzer, Mark Davison, Nicole Lurie and Bryan Dowd for their helpful comments on an earlier version of this paper.

Editor's Note

This manuscript is the first of a four-part series. Subsequent installments will appear in the next three issues of the Journal of Clinical Epidemiology.

REFERENCES

4. Miller GA. The magical number seven plus or minus two: some limits on our capacity to process information. Psychol Rev 1956; 63: 81-97.