

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.

Values preferences Preference weights Social preferences Utility measurement Health-state preferences Health status measurement

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-

*Reprint requests should be addressed to: Debra Froberg, Ph.D., Division of Human Development and Nutrition, University of Minnesota, School of Public Health, Box 197 UMHC 420 Delaware Street S.E., Minneapolis, MN 55455, U.S.A.

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

Mobility	Pain	Emotional well-being
No limitations	No pain	Not depressed
Walks with a limp	Mild pain	Slightly depressed
Uses a crutch or aid	Moderate pain	Moderately depressed
Does not walk	Severe pain	Very depressed

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, each to be experienced for a specified period of time, and asks which alternative is preferred. 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 8 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.

*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.

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.

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 is efficient in that it requires fewer respondent judgments and permits an analysis of the separate effects of each attribute. Yet, Veit and Ware

[8] point out that, like 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 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, ψ . 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

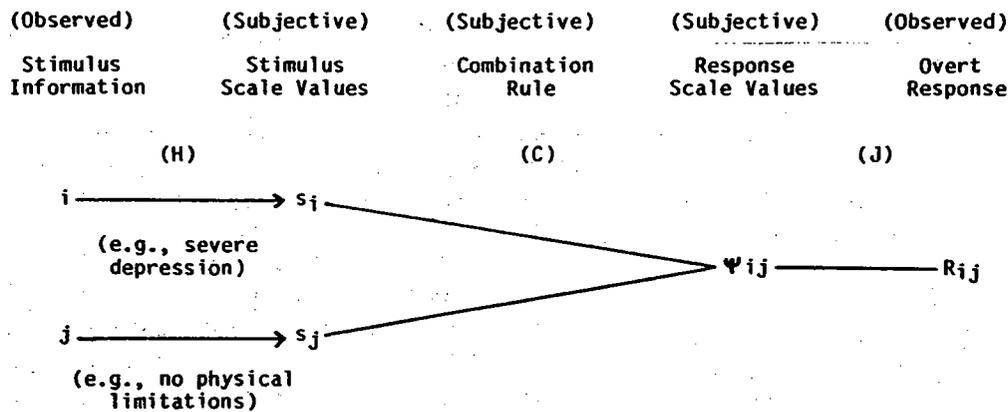


Fig. 1. Outline of subjective processes. (Veit *et al.* [11])

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 (ψ_{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

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

Adapted from Veit and Ware [8].

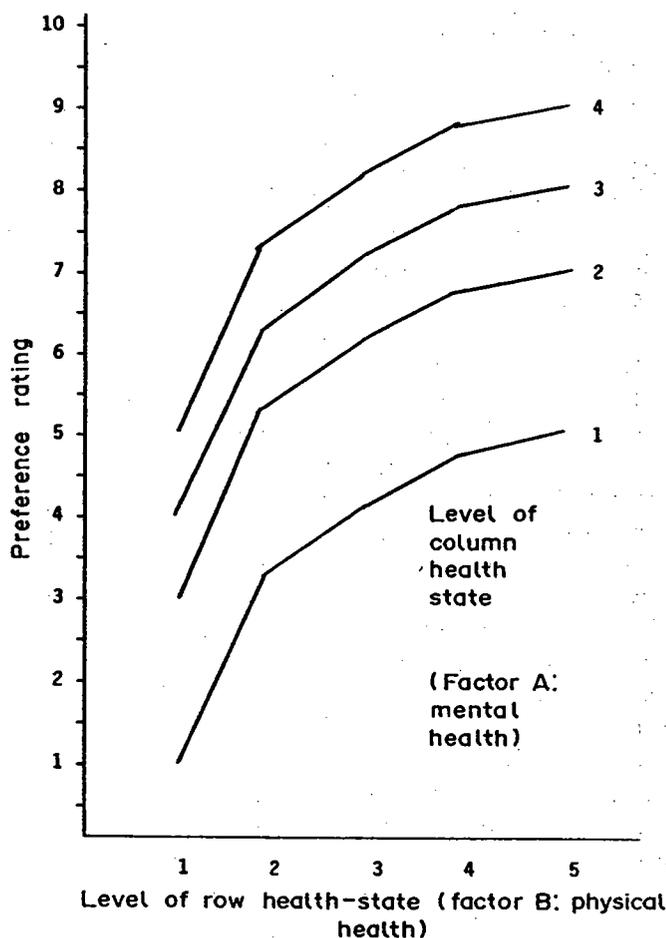


Fig. 2. Preference ratings of Factor B for each level of Factor A using data from Table 2. Adapted from Veit and Ware [8].

The same logic applies in the validation of the multilinear models: if the interaction effects can be attributed to cross-products of main effects, then both the model and scale values are simultaneously validated.

The shortcomings of functional measurement are primarily logistical in nature. First, where there are many attributes and levels within attributes, the number of multiattribute judgments required to achieve a complete factorial design may be prohibitive. However, this problem is not insurmountable since sophisticated "fractional" designs may be employed that produce the necessary information from a smaller number of multiattribute judgments. Second, application of these techniques requires technical expertise in the area of experimental design and analysis-of-variance, particularly if nonadditive models are involved. Interpretation of analysis-of-variance is not always straightforward. For example, the analysis may lack power to detect small interactions, or conversely, statistically significant four-, five-, or six-way

interactions may be impossible to meaningfully interpret.

A third 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].

Multiple regression. Because in many contexts, attempts to infer the parameters of statistical models have relied on multiple regression rather than the more sophisticated functional measurement procedures, we will say a few words about this approach. Regression procedures have been used to obtain some understanding of the combination rule (*C* in Fig. 1). This approach involves asking judges to evaluate a set of multiattribute health states, then estimating the subjective weights and scale values of a simple utility model (usually the

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

1. Feinstein AR. *Clinimetrics*. New Haven: Yale University Press; 1987.
2. Bush JW. Relative preference versus relative frequencies in health-related quality of life evaluation. In: Wenger N, Mattson M, Furberg C, Elinson J, Eds. *Assessment of Quality of Life in Clinical Trials of Cardiovascular Therapies*, New York: Le Jacq; 1984.
3. Torrance GW, Boyle MH, Horwood SP. Application of multi-attribute utility theory to measure social preferences for health states. *Oper Res* 1982; 30: 1043-1069.
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.
5. Bombardier C, Tugwell P, Sinclair AJ. Preference for endpoint measures in clinical trials: results of structured workshops. *J Rheumatol* 1982; 9: 793-800.
6. Spitzer WO, Dobson AJ, Hall J, Chesterman E, Levi J, Sheperd R, Battista RN, Catchlove BR. Measuring the quality of life of cancer patients: a concise QL-index for use by physicians. *J Chron Dis* 1981; 34: 585-597.
7. Boyle MH, Torrance GW. Developing multiattribute health indexes. *Med Care* 1984; 22: 1045-1057.
8. Veit CT, Ware Jr JE. Measuring health and health-care outcomes: issues and recommendations. In: Kane RL, Kane RA, Eds. *Values and Long Term Care*. Lexington, Mass.: Lexington Books; 1982.
9. Cadman D, Goldsmith C. Construction of social value or utility-based health indices: the usefulness of factorial experimental design plans. *J Chron Dis* 1986; 39: 643-651.
10. Fischer GW. Utility models for multiple objective decisions: do they accurately represent human preferences? *Decis Sci* 1979; 10: 451-479.
11. Veit CT, Rose BJ, Ware Jr JE. Effects of physical and mental health on health-state preferences. *Med Care* 1982; 20: 386-401.
12. Patrick DL, Bush JW, Chen MM. Toward an operational definition of health. *J Health Soc Behav* 1973a; 14: 6-23.
13. Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. *J Chron Dis* 1978; 7: 347-358.
14. Keeney RL, Raiffa H. *Decisions with Multiple Objectives*. New York: John Wiley; 1976.
15. Anderson NH. Integration theory and attitude change. *Psychol Rev* 1971; 78: 171-206.
16. Anderson NH. Algebraic models in perception. In: Carterette EC, Friedman MP, Ed. *Handbook of Perception*. New York: Academic Press; 1974: Vol. 2.
17. Llewellyn-Thomas HA, Sutherland HJ, Ciampi A, Etezadi-Amoli J, Boyd NF, Till JE. The assessment of values in laryngeal cancer: reliability of measurement methods. *J Chron Dis* 1984; 37: 283-291.
18. Lyness KS, Cornelius ET. A comparison of holistic and decomposed judgment strategies in a performance rating simulation. *Organ Behav Hum Perf* 1982; 29: 21-38.
19. Wiggins N, Hoffman PJ. Three models of clinical judgment. *J Abnorm Psychol* 1968; 73: 70-77.
20. Huber GP, Sahney VK, Ford DL. A study of subjective evaluation models. *Behav Sci* 1969; 14: 483-489.
21. Hoepfl R, Huber G. A study of self-explicated utility models. *Behav Sci* 1970; 15: 408-414.
22. Birnbaum MH. The devil rides again: correlations as an index of fit. *Psychol Bull* 1973; 79: 239-242.
23. Birnbaum MH. Reply to devil's advocate; don't confound model testing and measurement. *Psychol Bull* 1974; 81: 854-859.
24. Arrow KJ. *Social Choice and Individual Values*. New York: John Wiley; 1951.
25. Keeney RL. A group preference axiomatization with cardinal utility. *Management Sci* 1976; 23: 140-145.
26. Resnick MD. *An Introduction to Decision Theory*. Minneapolis: University of Minnesota Press; 1987.

METHODOLOGY FOR MEASURING HEALTH-STATE PREFERENCES—II: SCALING METHODS

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—This paper begins with a discussion of measurement principles relevant to determining health-state preferences. Six scaling methods are described and evaluated on the basis of their reliability, validity, and feasibility. They are the standard gamble, time trade-off, rating scale, magnitude estimation, equivalence, and willingness-to-pay methods. Reliability coefficients for most methods are acceptable although the low coefficients for measurements taken a year apart suggest that preferences change over time. Convergent validity among methods has been supported in some but not all studies, and there are limited data supporting hypothetical relationships between preferences and other variables. The category ratings method is easiest to administer and appears to yield valid scale values; thus, it is recommended for large-sample studies. However, decision-oriented methods, particularly the time trade-off and standard gamble, may be more effective in small-scale investigations and individual decision making.

Values preferences Preference weights Social preferences Utility measurement Health-state preferences Health status measurement

INTRODUCTION

After deciding whether to use a holistic or decomposed strategy to gather data on health-state preferences as discussed in Part I, the investigator faces a choice among scaling methods. This choice has been the subject of a great deal of attention in the literature, with different investigators presenting arguments for the superiority of the standard gamble, time trade-off, magnitude estimation, category ratings, equivalence, and willingness-to-pay techniques. Our description and comparison of these methods builds on a long tradition of psychometric research.

Scientists engaged in the study of psychophysics have provided encouraging results supporting the validity of subjective judgments in general. Psychophysics is concerned with the way in which people perceive and make judgments about physical phenomena such as the brightness of lights or loudness of sounds. Since about the mid-1800s, scientists have been interested in establishing mathematical relationships between stimulus intensity and sensation. They have discovered that this relationship is not always linear (e.g. doubling the intensity of a light will not cause people to report it as twice as bright). Nonetheless, humans can make consistent, numerical estimates of sensory stimuli. The exact form of the relationship varies from one sensation to another, described by an equation with a different power function exponent for each type of stimulus: $R = KS^b$ where R is the response, S is the level of the stimulus and b is an exponent that typically falls within the range of 0.3–1.7.† Research validating the power law has led to the conclusion that people can make remarkably consistent subjective

*Reprint requests should be addressed to: Debra Froberg, Ph.D., Division of Human Development and Nutrition, University of Minnesota, School of Public Health, Box 197 UMC 420 Delaware Street S.E., Minneapolis, MN 55455, U.S.A.

†It happens that for judging line lengths, the exponent is unity, which means that the relationship between stimulus and response is linear. This convenient result underlies the interpretation of visual analogue scaling methods discussed later.

judgments, even when those judgments are abstract [1].

Psychophysical methods have been adapted for use in measuring subjective judgments for which there is no physical scale, including preferences and values. This is the field of psychometrics, to which we now turn for some basic concepts and definitions.

MEASUREMENT CONCEPTS AND DEFINITIONS

In simple terms, the problem addressed in this paper is one of quantifying or measuring preferences for health states. Four sets of distinctions are relevant to this discussion: (1) scaling stimuli as opposed to scaling persons, (2) scaling verifiable vs nonverifiable stimuli, (3) levels of measurement produced by various scaling methods, and (4) direct vs indirect scaling methods.

Scaling stimuli vs persons

For purposes relevant to this paper, when we ask people to rate the desirability of a set of health states, we are engaged in a stimulus-scaling task, the stimuli being the health states. This is distinct from the more familiar measurement situation in which the goal is to scale people. An example of the latter is when we assign numbers to people based on their responses to an instrument that measures health status. To better understand this distinction, the question could be asked: are we interested in comparing people by identifying their location on a continuum, or are we comparing something else on a continuum, namely, health states?

The distinction between scaling stimuli and scaling persons is important for two reasons. First, it has implications for selecting an appropriate scaling technique. For example, Likert methods are generally used for scaling persons, magnitude estimation is for scaling stimuli, and Guttman scales accomplish both. Second, it has implications for the way in which variability in preferences is handled. In stimulus scaling, the objective is usually to obtain consensus among judges as to the scale values for each stimulus, whereas the objective in scaling persons is to discriminate among persons by spreading them out on a continuum [2].

Verifiable vs nonverifiable stimuli

The second important distinction is whether or not the subjective scale can be compared to some external standard of accuracy. In a typical

psychophysics experiment where subjects are asked to adjust one light so it appears twice as bright as another light, the ratio of perceived brightness can be compared with the ratio of physical magnitudes of illumination. In a study of health-state preferences, there is no factual standard against which to compare subjects' responses. (Comparing subjects' stated preferences with their behavior would provide interesting information but lack of correspondence between the two would not necessarily mean the stated preferences are "incorrect".) The importance of this distinction will become clearer later when we discuss the validation of scaling methods. The validation process for health preferences scaling methods involves the incremental accumulation of evidence rather than any one definitive comparison.

Level of measurement

A third set of distinctions concerns the level of measurement produced by various scaling methods. Measurement scales can be classified as (1) categorical or nominal, (2) ordinal, (3) interval, or (4) ratio [3]. These categories represent different uses made of numbers and the legitimacy of performing particular classes of mathematical procedures. *Categorical* measurement is not of interest in this paper since it is more accurately a means of identification than of quantification. For example, males could be assigned a number of 1 and females a number of 2, but these numbers are not intended to have quantitative meaning. An *ordinal* scale, the most primitive form of measurement, is one in which a set of objects (e.g. health states) is rank-ordered, but there is no indication of how much of the attribute (e.g. desirability) each object possesses nor how far apart the objects are with respect to the attribute. An *interval* scale does provide information on how far apart the health states are as well as their rank order, but it does not indicate the absolute magnitude of desirability for any health states. A *ratio* scale is achieved when, in addition to knowing the rank order of the health states and how far apart they are, it is possible to know the distance from a rational zero for at least one health state. This latter characteristic enables the absolute amount of desirability to be determined for all health states.

Scaling methods differ in the level of measurement they achieve. It is important to know what level a particular scaling method produces because the higher the level of measurement, the

more forms of mathematical treatment can be applied to the data. The ratio scale is susceptible to the fundamental operations of algebra: addition, subtraction, division and multiplication. In addition, a ratio scale remains invariant over all transformations where the scale is multiplied by a constant. This means that the scale remains essentially the same when it is expressed in different units (e.g. feet rather than inches). The potential disadvantage of having only an interval scale is that algebraic operations can only be performed on intervals and not on scale values, so it cannot be said, for example, that one health state is twice as desirable as another. However, for most practical purposes an interval scale is sufficient. Most powerful methods of statistical analysis require only interval scales. Health status indexes depend upon values being measured on an interval scale [4], and for cost-effectiveness analysis, a unique solution results if interval scale numbers are used [5]. Ordinal scales provide only meager information and none of the fundamental operations of algebra may be applied. Unfortunately, in practice, scale properties often go untested and ordinal data are treated as if they were interval data.

Direct vs indirect scaling

In direct scaling, respondents are instructed to make judgments at a certain level of measurement and the resulting data are treated as such. For example, respondents may be asked to perform a ratio-level scaling task such as assigning a number representing the absolute magnitude of disability to each of a series of health states. Conversely, in indirect scaling, respondents are instructed to make their judgments at a certain level of measurement, and the data are later converted to a different level by the investigator. For example, in the method of paired comparisons, all possible pairs of health states are presented to respondents, and they need only indicate which of the two states represents greater disability (an ordinal judgment). In order to convert these ordinal-level judgments to interval-level data, the experimenter must apply a set of theoretical assumptions based on the variability of subjects' responses. One such set of assumptions, known as Thurstone's Law of Comparative Judgment, is based on the idea that stimulus differences which are detected equally often are subjectively equal [6].

Direct scaling methods can be thought of as methods in which the step between the raw data and final scale is as short as possible. Typically, when respondents provide interval- or ratio-level data, the investigator can derive scale values through relatively simple computations such as averaging across respondents [7]. This direct approach to scaling used to be considered by psychometricians as too "subjective"; however, recent evidence supports the validity of direct scaling methods, and their ease of use and simplicity have led to their exclusive use in health preference measurement. For our purposes, the important distinction between direct and indirect scaling models lies in their assumptions.

Direct scaling models assume that: (1) the subject is capable of directly generating an interval or ratio scale (2) there is some error in the judgments made by one person on one occasion but error can be reduced by averaging judgments over subjects, i.e. subjects are replicates of one another.

Since all scaling models used in the health preference literature have used direct scaling, they have taken seriously the subject's ability to generate interval and ratio scales directly. It should be noted that this is an *assumption* underlying direct scaling methods. Whether or not the scale values resulting from these methods truly are at the interval level of measurement is an empirical question which can and should be tested. The functional measurement design presented in Part I provides a means for testing the interval property of scale values through testing a model of information processing. If the data support the model, both the level of measurement and model form are validated simultaneously.

DESCRIPTION OF SCALING METHODS

Three scaling methods used in studies of health-state preferences require subjects to respond in terms of an interval scale: the standard gamble, time trade-off and category ratings [8]. The other scaling methods (magnitude estimation, equivalence, and willingness-to-pay) require ratio-level responses. Each of these techniques will be briefly described before comparing their relative merits.

The *standard gamble* is the classical method of measuring preferences originating in the field of decision theory.* First presented by von Neumann and Morgenstern [10], it is based on

*It is similar in concept to the traditional psychometric methods of fractionation [9].

the axioms of utility theory and incorporates a conceptual framework for examining decision making under uncertainty. The essence of the technique is a choice posed to the respondent between a certain outcome and a gamble. Figure 1 illustrates the standard gamble for a chronic health state preferred to death. The choice is usually presented to the respondent as one of accepting or not accepting a treatment. The treatment (alternative 1) is a gamble with two possible outcomes:

"Either the patient is returned to normal health and lives for an additional t years (probability p), or the patient dies immediately (probability $1 - p$). Alternative 2 has the certain outcome of the chronic state i [e.g., hospital dialysis] for life (t years). Probability p is varied until the respondent is indifferent between the two alternatives, at which point the required preference value for state i is simply p ; that is, $h_i = p$ " [8, p. 20].

An intuitive way of understanding the standard gamble is the following: if state i in Fig. 1 is very undesirable (say, complete paralysis) then a respondent will be willing to take a treatment gamble even if the probability (p) of returning to full health is rather low (say, 0.30). Thus, the scale value for complete paralysis is also low. Variations in the standard gamble techniques are possible if the investigator is interested in states worse than death or temporary health states.

Figure 2 shows how the standard gamble is applied to states worse than death. Here the certain alternative is death, whereas the gamble alternatives are healthy (with probability p) or state i (with probability $1 - p$). A common way of presenting this is to ask the subject to imagine that he/she has a terminal disease which will lead to death if untreated. If treated, there is a probability p that the disease will be cured, but a probability $1 - p$ that the subject will fall into chronic state i , the state worse than death. The probability p is varied until the subject is indifferent between the two alternatives, at which point the preference value for state i is

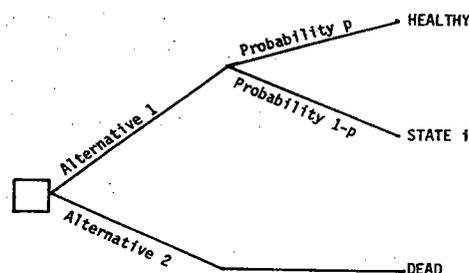


Fig. 2. Standard gamble for a chronic health state considered worse than death. (From Torrance [8].)

given by $h_i = -p/(1 - p)$. Intuitively, this means that if state i is very undesirable (say, a chronic vegetative state), then a respondent would not choose the gamble unless the probability of returning to a healthy state were very high. It also means that states worse than death are represented by negative numbers.

The standard gamble can be applied to temporary states as shown in Fig. 3. Here the certain alternative is state i , the state to be measured, just like in the chronic health state example shown in Fig. 1. The difference between Figs 1 and 3 is that the gamble in Fig. 3 replaces "dead" with the worst temporary state. The formula used to compute the value of state i is $h_i = p + (1 - p)h_j$ [8].

Regardless of the variant used, the standard gamble always poses a choice between a gamble and a certain outcome, where the certain outcome is intermediate in desirability between the best and worst gamble outcomes. To make it easier for subjects to think in terms of probabilities, the standard gamble is often presented with the aid of a probability wheel. This is a disk with two moveable, different-colored sections which can be adjusted to represent the probabilities of the two gamble alternatives, p and $1 - p$. In addition, rather than requiring respondents to decide immediately upon a probability, investigators generally use a "back and forth" technique, beginning by asking if the respondent would take the treatment at probability levels of 1.0 or 0.0. The investigator progressively

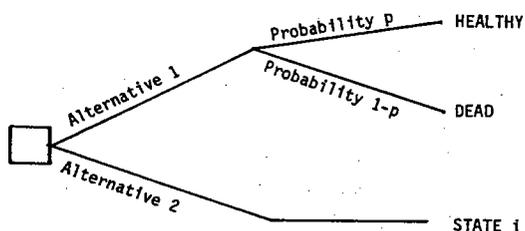


Fig. 1. Standard gamble for a chronic health-state preferred to death. (From Torrance [8].)

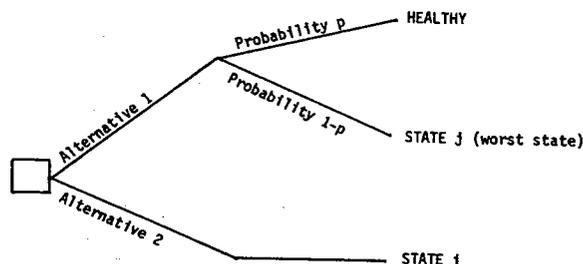


Fig. 3. Standard gamble for a temporary health state. (From Torrance [8].)

narrows the probability range until the respondent is able to choose a specific probability.

As the preceding discussion shows, the standard gamble is complex and not intuitively obvious to most respondents. The *time trade-off* method was developed by Torrance and his colleagues [11] specifically for use in health research as a simple-to-administer alternative to the standard gamble. Like the standard gamble, it presents the respondent with a choice. However, in the time trade-off technique the respondent is asked to choose between two alternatives of certainty rather than between a certain outcome and a gamble. The technique asks the respondent how much time (years of life) he or she would be willing to give up to be in a healthier state compared with a less healthy one. Figure 4 shows the time trade-off method for chronic states considered better than death. Torrance [8] describes the procedure as follows:

The subject is offered two alternatives—alternative 1: state i for t (life expectancy of an individual with the chronic condition) followed by death; and alternative 2: healthy for time $X < t$ followed by death. Time X is varied until the respondent is indifferent between the two alternatives, at which point the required preference value for state i is given by $h_i = X/t$ [8, p. 23].

The time trade-off method can be altered to apply to health states considered worse than death and for temporary states. When temporary states are measured relative to each other, state i (the state being scaled) can be compared to any other state (j) as long as state j is considered worse than state i . Again, to make the task easier, a procedure of starting at the extremes and converging toward the middle is used to help respondents decide upon a time, X . Torrance [12] has developed visual aids consisting of a laminated cardboard with sliders, and changeable scales and health states.

Originating in psychometrics, the *rating scale* consists of a line on a page with clearly defined endpoints or anchors. It requires that the respondent identify the best and worst health states to use as anchors. (In practice the anchors are usually labeled "death" and "perfect

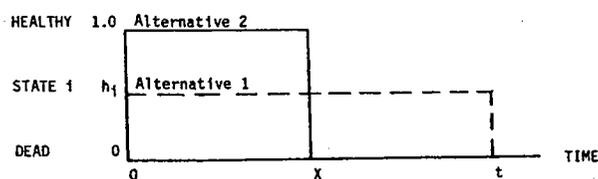


Fig. 4. Time trade-off for a chronic health-state preferred to death. (From Torrance [8].)

health".) The respondent then rates the desirability of each health state by placing it at some point on the line between the anchors. To achieve an interval scale, respondents must be instructed to place the health states on the line such that the intervals between the placements reflect the differences they perceive between the health states. A commonly used variation of the rating scale method is the method of equal appearing intervals, or category ratings. In this procedure, respondents sort the health states into a specified number of categories (often 10), assuming equal changes in preference between adjacent categories.

Visual aids may be used with either form of the rating scale. A thermometer with a scale from 0 to 100 on a felt background has been used, along which respondents place foam sticks labelled with the health states. In the equal-appearing-intervals method, respondents may actually sort cards labelled with health states into piles, or they may simply assign a category number to each health state. The rating scale is the most frequently used method for measuring health-state preferences. It can be used for scaling either chronic or acute states as well as states worse than death [12].

Magnitude estimation is a scaling method proposed by Stevens [13] to overcome what he saw as limitations of the category ratings method; namely, the lack of ratio-level measurement and the supposed tendency for subjects to use categories equally often. Using magnitude estimation, the respondent is given a standard health state and asked to provide a number or ratio indicating how much better or worse each of the other states is compared with the standard. For example, Kaplan *et al.*'s [14] instructions were as follows:

"Let's give the first case the number 10. Now assign numbers to the other cases using the number 10 as your guide. For example, if a case seems 10 times as desirable as the first case you would use a number 10 times as large or 100. If it seems one-fifth as desirable you would use the number 2 and so forth. Use fractions, whole numbers or decimals, but make each assignment in relation to the desirability of the first case, as you see it" [14, p. 525].

Studies using this method have been inconsistent in the selection of a standard health state. In three studies, the standard was taken from the end of the scale, defined as the least ill state, the healthiest state, or the absence of discomfort or dysfunction [15–17]; whereas in another study the standard was taken from the middle

of the scale [14]. These studies also differed in the direction of the scale, with some defining 0 as the least desirable health state and others defining it as the most desirable.

Two other scaling methods, equivalence and willingness to pay, have been used less frequently but deserve mention. *Equivalence* is an adaptation of the method described in psychometric literature as the method of adjustment or equivalent stimuli. It has been applied in various forms [e.g. 16, 18], but the common underlying task for the respondent is to decide how many people in health state B are equivalent to a specified number of people in health state A. For example, in one study [16], respondents were instructed as follows:

The first group contains 100 people in a state of maximum health (standard). Persons in the second group are in the state of health lower than the standard [specified]... How many people in this state of health do I consider equivalent to the 100 people of the same age in the standard group?... You may use any number equal to or greater than 100 [16, p. 236].

The equivalence technique is conceptually similar to magnitude estimation. In fact, when Rosser and Kind [15] used magnitude estimation, they attempted to clarify the implications of the method to respondents in terms of the equivalence method.

Thompson [19] recommends the *willingness-to-pay* technique as a means of measuring health preferences. This technique has been used in cost/benefit and cost/effectiveness analyses to quantify programs that are difficult to value in monetary terms. Although its use in assigning values to health states has been limited to date, it has been used extensively in valuing changes in the risks of dying [20]. The willingness-to-pay method, as applied by Thompson [19], consisted of the question: what percent of your family's (i.e. household) income would you be willing to pay on a regular basis for a complete cure of arthritis? Respondents were instructed to assume that a complete cure existed, that their insurance would not cover it, and that they would have to pay for it. The responses were expressed as proportion of income. In an earlier study, Thompson *et al.* [21] also asked respondents how much (in dollars) they would be willing to pay each week to get rid of their arthritis. However, he concluded that the dollar amount is less useful than proportion of income because it had far fewer associations with independent variables and is influenced by income level.

EVALUATION OF SCALING METHODS

Before we compare the performance of scaling methods on the basis of their reliability, validity, and feasibility, it may help to provide some background on three of the methods that have long histories of use in other disciplines. The standard gamble, rooted in decision theory, and category ratings and magnitude estimation, rooted in psychometrics, have relatively long histories of use even though their application to health-state preference is a rather recent development.

Decision theorists have historically favored the standard gamble because it is built on a set of fundamental axioms underlying the expected utility model and it forces the respondent to make preference judgments under conditions of uncertainty [22, 23]. Also, the standard gamble has been said to yield an interval scale [24], although such claims appear to be definitional rather than empirically demonstrated [25].

While the standard gamble has been viewed by some as the criterion scaling method due to its theoretical grounding in expected utility theory, some decision theorists have turned to other methods because the standard gamble is so difficult to explain to respondents. Further, recent evidence suggests that people exhibit patterns of preference that are incompatible with expected utility theory. For example, Llewellyn-Thomas *et al.* [26] found that changes in the gamble outcomes significantly influenced reported values for health states, a finding that both contradicts expected utility theory and indicates that the standard gamble is internally inconsistent. Shoemaker [27] presents extensive evidence that people violate the axioms of EUT. At the individual level, EU maximization is more the exception than the rule, at least for the types of decision tasks examined. These theoretical developments raise questions concerning the validity of the standard gamble technique.

In particular, utilities derived from the standard gamble may be biased by risk aversion. Economists generally accept the hypothesis that individuals are risk averse when evaluating a sure gain against a gamble with an equal or higher expected gain. However, psychological research indicates that when people are faced with a sure loss vs a gamble with a substantial probability of an even greater loss, they are often risk-seeking and choose the gamble. Putting these two pieces together, Kahneman

and Tversky [28] studied risky prospects that involved both positive and negative outcomes. The standard gamble, with a certain health state evaluated against a gamble with some probability of perfect health and some probability of death, is an example of a risky prospect with both positive and negative outcomes. Kahneman and Tversky [28] found that the pleasure of a gain is much less intense than the pain of a similar-sized loss. This finding suggests that people will usually choose to remain in a less-than-perfect health state rather than risk ending up sicker or dead. In particular, a health state would have to be extremely undesirable before a person would accept an operation with even a moderate risk of death. This conservatism with respect to risk taking would have a tendency to inflate utilities derived from the standard gamble relative to other scaling methods that do not involve gambles.

Much research in psychometrics has centered on the debate between category scaling and magnitude estimation. Category scaling developed out of early work in psychophysics, the study of mathematical functions relating physical intensities to internal sensations. This tradition stood for nearly 100 years (from 1860 to 1960) until it was challenged by Stevens [29–31], who contended that category scaling methods did not produce linear (interval) response scales. Stevens claimed that magnitude estimation was a superior scaling method due to its dependence on direct estimation of subjective ratios. In psychometrics today, while there are many advocates of magnitude estimation, category scaling continues to be most frequently used in applied areas. Stevens' argument in favor of magnitude examination is intuitively appealing, but he has failed to produce empirical evidence for its superiority over category scaling [4]. Experiments using functional measurement as a means of testing for equal intervals have shown that category ratings meet this empirical criterion while magnitude estimation does not [14, 32, 33]. Moreover, Kaplan and Ernst [4] demonstrated that a supposed bias inherent in category ratings, the distribution effect, does not necessarily occur. In their study, when subjects

rated health-state descriptions, they did not attempt to use all categories equally frequently.

The remaining scaling methods need less introduction because their histories are shorter. The time trade-off technique was recently developed by Torrance expressly for the scaling of health preferences. It was designed to produce the same results as the standard gamble at less cost and with less burden on the respondent. Willingness-to-pay has been applied in a number of cost-effectiveness analyses, but more often to measure the utility of reducing one's risk of dying than to measure preferences for various states of morbidity. The equivalence method may be viewed as an alternate form of magnitude estimation, and has been used only a few times in studies of health preferences.

Reliability

A measure is reliable if it is relatively free of measurement error. Reliability concerns the extent to which a scaling method produces consistent results. With respect to the scaling of health states, reliability can be assessed in three ways: intra-rater reliability refers to a single rater's consistency when an item is presented more than once; test-retest reliability refers to stability of scale values over short periods of time; and inter-rater reliability is consistency among judges regarding scale values.* Table 1 presents available data on each type of reliability for the different scaling methods. The most obvious observation is that the table has much missing data. Data on all three types of reliability are available only for rating scales.

In general, intra-rater reliability is acceptable for all scaling methods for which these data are available. Test-retest reliability coefficients up to 6 weeks are also satisfactory with the possible exception of 0.63, the lower range value for the time trade-off method at 6 weeks. Interpretation of the low test-retest reliability for measurements taken a year apart is ambiguous; the low coefficients probably reflect true preference changes as well as measurement error. Inter-rater reliability appears to be acceptable except for the rather low coefficient of 0.60 reported by Patrick *et al.* [16] for the equivalence method. Overall, these data are encouraging, but the gaps in the table indicate a need for further research. Also, comparisons among the studies are limited by the fact that a frequently used statistic, the Pearson Product Moment Correlation Coefficient, is dependent upon variability across subjects. Thus, correlations from studies

*Internal consistency reliability, or the consistency in response from item to item within a scaling task, is not applicable to the scaling of health states. There is no reason to expect high intercorrelations among the stimuli nor would they be desirable. Internal consistency is important in situations where a series of items are used to scale people—not stimuli—on a particular dimension.

Table 1. Reliability of scaling methods^a

Reliability	SG	TTO	RS	ME	EQ	WTP
Intra-rater reliability	0.77 [38]	0.77-0.88 [38, 52]	0.70-0.94 [16, 49, 52]	0.74-0.83 [16]		
Intra-rater agreement (%)				97.2% [15]		
Test-retest reliability						
1-week or less	0.80 [51]	0.87 [51]	0.77 [51]			
4-week		0.81 [42]				
6-week		0.63-0.80 [50]				
1-year	0.53 [38]	0.62 [38]	0.49 [38]			0.25 [21]
Inter-rater reliability			0.75-0.77 [16]	0.75-0.79 [16]	0.60 [16]	
Inter-rater agreement (%)				88% [15]		

SG = standard gamble; TTO = time trade-off; RS = rating scale; ME = magnitude estimation;

EQ = equivalence; WTP = willingness-to-pay.

^aAll are correlations unless otherwise indicated.

using different subjects and sample sizes are not directly comparable [34].

This table does not include test-retest measurements taken before and after therapy, or before and after some other event that would be likely to change one's preferences. For example, Christensen-Szalanski [35] reports reliability coefficients ranging from 0.37 to 0.59 for two measurements of women's preferences for anesthesia during childbirth. The first measure was taken during labor and the second one at one month postpartum. Although the stability of these two measurements was rather low, there was high concordance between preferences one month prior to delivery and one month after delivery. Not surprisingly, preferences for anesthesia were more positive during labor than at the other two times. This study highlights the problem of differences between current and long-term values.

In a second study, two sets of category ratings were obtained, one before patients began chemotherapy and the other 6 weeks later after treatment. The correlation between the two sets of ratings was only 0.17. However, in the same study under the same conditions, the standard gamble produced more stable preference values, with coefficients ranging from 0.48 to 0.59 [36].

In contrast to these two studies, Llewellyn-Thomas *et al.* [37] found that patients' values were uninfluenced by a deterioration in their own clinical state brought on by radiotherapy. Despite the fact that laryngeal cancer patients experienced a deterioration in their voice

quality, reliability of their preferences for aspects of voice quality remained stable throughout therapy. Discrepant findings among these three studies imply a need to further examine the causes of preference shifts.

Validity

A scaling method is valid if it accurately measures what it is intended to measure. Validity is generally thought to be of three types: content, criterion, and construct. Construct validation is the most comprehensive, and some measurement experts view it as encompassing the other two types. Applied to health-state preferences, content validity refers to the adequacy of the health-state descriptions in representing health status. Content validity is achieved by careful selection of attributes (discussed in Part I) and presentation of sufficient detail. Studies of health-state preferences differ widely in the content of health-state descriptions, and unfortunately, content validation is rarely discussed. (However, it is discussed with respect to particular health status measures, such as the Sickness Impact Profile.) We will discuss some aspects of content validity in Part III, under the heading of situation-specific variables. Strictly speaking, criterion-related validity does not apply to health-state preferences since there exists no criterion embodying individuals' "true" preferences, nor are we attempting to predict some future behavior.

In scaling preferences, we are concerned with an abstract variable or "construct" rather

than an observable one. To define this abstract variable and determine what a particular scaling method actually measures requires methods of construct validation [3]. Many approaches to construct validation are possible, two of which have been taken in the validation of health preference scaling methods: (1) examining the degree to which results of different scaling methods converge, and (2) examining the degree to which predicted relationships between preferences and other variables are empirically supported. Considerably more work has been done using the first approach than the second.

Convergence of methods. Studies comparing scaling methods have either examined the functional relationships between the methods or compared the mean scale values derived from each method. Two studies have compared category ratings and magnitude estimation. While an early study [16] found that category ratings and magnitude estimation were linearly related, a later study [14] found a logarithmic relationship. The later study is more consistent with related psychometric research both in methodology and findings. Kaplan and his colleagues [14] concluded that, because scale values derived from magnitude estimation were compressed to the lower extreme of the scale near death, magnitude estimation is not a valid scaling method for health preferences. However, a recent study contests Kaplan *et al.*'s conclusion, claiming that Kaplan chose an inappropriate zero point. According to Haig *et al.* [17], the correct zero point should be the absence of dysfunction and discomfort, not death. A possible reason for so many of Kaplan's scale values clustering at zero (death) is that using death as an anchor created a floor effect, making it impossible to rate states as worse than death. When Haig and his colleagues inverted the scale and assigned 0 to the absence of dysfunction and discomfort, they found linear relationships between their magnitude scale and the category ratings reported by Bush *et al.* in an earlier study. In general, studies in which the "standard" for magnitude estimation is a perfectly well state show no differences in scale values obtained with category and magnitude methods [4].

Three studies have compared the standard gamble, time trade-off, and rating scale. Torrance [38] viewed the standard gamble as the criterion technique, arguing that the standard gamble is valid by definition since it is based directly on intuitively appealing axioms of util-

ity theory for decisions made under uncertainty. (Note, however, that Shoemaker [27] presents considerable evidence that people do not act in accordance with these axioms.) He found a correlation of 0.65 between the time trade-off and standard gamble and a correlation of 0.36 between category ratings and the standard gamble. He also reported that individual and population mean values of the standard gamble and time trade-off appeared to be equivalent while category ratings were clearly different.

Wolfson *et al.* [39] arrived at a different conclusion after comparing the same three scaling methods. They found that values obtained for the standard gamble were consistently higher than those obtained for category ratings or time trade-off. The latter two were more similar than either was to the standard gamble. The authors speculate that scale values from the standard gamble are contaminated by an "aversion to gambling". Despite their contradictory findings both Torrance and Wolfson *et al.* recommend the use of the time trade-off method because it appears valid and is easier to administer than the standard gamble.

Read *et al.* [34] found moderately high correlations between the standard gamble, time trade-off, and category rating methods ($r = 0.56-0.65$) for both single-attribute and multi-attribute health states. However, the standard gamble generated consistently higher preference scores than the other two methods. In addition, for multiattribute health states there was a significant interaction between two attributes, angina severity and length of survival, using category scaling, but not using the standard gamble. These authors stress that high correlations among scaling methods do not guarantee that the methods produce equivalent ratings. Two additional studies compared only the standard gamble and category ratings. Both found standard gamble values to be significantly higher than category rating values [36, 40], and one also reported nonsignificant correlations between the two methods [36].

One study [41] compared the time trade-off, category ratings, and a third approach called direct questioning of objectives. (This method involved the use of a category scale to measure the patient's ability to achieve objectives of importance to him or her.) When patients rated their present health states using each method, the mean values were almost identical. In contrast to these convergent findings, Churchill *et al.* [42] found only a low correlation (0.22)

between the time trade-off method and a visual analogue rating scale.

Rosser and Kind [15] validated magnitude estimation by comparing it with fractionation and multiplication methods. Fractionation requires that the subject identify a state that is half as severe as a "standard" health state and then a third state half as severe as the second. The multiplication method requires the subject to select a state twice as severe as the standard, and then a third state twice as severe as the second. They found that nine out of ten subjects produced consistent responses across all three methods.

Only one study has compared the equivalence method with other scaling methods. Miles [43] found that differences between category ratings and equivalence were nonsignificant in each of 12 comparisons. No studies have directly compared willingness-to-pay to other methods, but Thompson [19] provides indirect evidence of a lack of convergence between it and the standard gamble. He conducted regressions of willingness-to-pay and maximum acceptable risk (derived from the standard gamble) on 31 other variables and found that different variables were associated with willingness-to-pay values than with standard gamble values. For example, pain was correlated with standard gamble values but not willingness-to-pay.

Table 2 summarizes the studies that have compared the results of different scaling methods. A "yes" in the table indicates that the investigators found at least one of three conditions: (1) a linear relationship between scaling methods, (2) a significant correlation between scaling methods (which doesn't necessarily imply a strict linear relationship) or (3) that the

mean values were not statistically different. Even using this liberal criterion, the table shows that these studies have produced mixed results. A substantial amount of convergence is evident, but no clear patterns emerge concerning which methods do and do not converge. Perhaps the most that can be concluded is that while correlations between methods are usually moderately high, the different methods do not necessarily produce equivalent scale values.

More research is needed to further explore the convergence of scaling methods, particularly the two that have not yet been studied, equivalence and willingness-to-pay. However, in the psychosocial measurement literature, it is generally accepted that although different scaling methods should produce the same rank ordering, they should not necessarily be expected to produce identical results. The exact scale values produced by different methods will differ because the methods ask respondents to perform different tasks, perhaps invoking entirely different cognitive processes. For example, magnitude estimation methods ask respondents to judge magnitudes while category ratings require the judgment of intervals. The underlying scale elicited by these methods depends on the task [44]. The task itself may influence such cognitive activities as attention to a particular stimulus, recall of past experiences, selection of reference points, and emotional reactions—all of which might influence one's evaluations of health outcomes [34].

Thompson's [19] study comparing the standard gamble and willingness-to-pay methods illustrates another way in which the task influences the response scale. Regression analysis showed that arthritis patients seemed to focus

Table 2. Convergence of scaling methods^a

Study	SG	TTO	RS	ME	EQ	WTP
Patrick <i>et al.</i> [16]			Yes	Yes	No	
Kaplan <i>et al.</i> [14]			No	No		
Haig <i>et al.</i> [17]			Yes	Yes		
Torrance [38]	Criterion	Yes	No			
Wolfson <i>et al.</i> [39]	No	Yes	Yes			
Read <i>et al.</i> [34]	Yes	Yes	Yes			
Llewellyn-Thomas <i>et al.</i> [40]	No		No			
O'Connor <i>et al.</i> [36]	No		No			
Detsky <i>et al.</i> [41]		Yes	Yes			
Churchill <i>et al.</i> [42]		No	No			
Miles [43]			Yes		Yes	

SG = standard gamble; TTO = time trade-off; RS = rating scale; ME = magnitude estimation; EQ = equivalence; WTP = willingness-to-pay.

^aA "yes" in the table indicates that investigators found at least one form of convergence: a linear relationship, a significant correlation, or mean values that were not significantly different.

on different aspects of their disease in responding to the two methods. For willingness-to-pay, the dominant health-related concern was for impairments in activities of daily living; for the standard gamble, it was pain. "It seems that people contemplating spending more money for arthritis care ask themselves how they could improve functionally. In pondering acceptable mortal risks, they are more strongly guided by their current levels of pain" [19, p. 394].

Selection of an appropriate scaling method thus depends upon the way in which the results will be used. In addition, further research elucidating relationships between the results of different scaling methods and other external criteria will enhance our understanding of what these scaling methods actually measure. The few existing studies of this nature are discussed in the next section.

Testing predictions. A few studies have tested hypothetical relationships between health-state preferences and other variables. Churchill *et al.* [42] asked end-stage renal patients to rate their own health using the time trade-off method, predicting that the mean scores would be highest for transplant patients, lowest for hospital hemodialysis patients, and intermediate for home/self-care hemodialysis and continuous ambulatory peritoneal dialysis patients. The results confirmed these predictions with time trade-off scores ranging from 0.43 (hospital dialysis) to 0.84 (transplantation).

Kind *et al.* [45] asked the question: to what extent are valuations of health states using magnitude estimation consistent with the values implied in court awards? They examined over 200 British court awards for damages in personal injury claims and found that the legal scale was significantly correlated (0.82) with the magnitude estimation scale.

Evidence supporting the validity of the willingness-to-pay method has been reported by Thompson *et al.* [21]. Consistent with their predictions, willingness-to-pay (as a proportion of income) was positively associated with the number of symptoms experienced by each patient and with such indices of health services utilization as the number of medicines taken and having had total knee replacements.

Christensen-Szalanski [35] found that women's preferences concerning the use of anesthesia during childbirth were significantly related to their decision to request anesthesia; however, the women did not request anesthesia as early in labor as their preferences indicated.

Feasibility

To be useful, scaling methods must be both economical and acceptable to respondents. The standard gamble and time trade-off are inherently expensive due to their reliance on a lengthy interview with well-trained interviewers using elaborate branching procedures. Further, because people find it difficult to work with probabilities and may also have an aversion to taking risks, they often do not give consistent and sensible answers to standard gamble questions even under standardized conditions [19]. This is particularly problematic in population studies with large numbers of subjects. However, the standard gamble is reportedly quite feasible in clinical situations where the physician or counselor can spend sufficient time with patients to carefully explain concepts of risk and uncertainty [46]. The standard gamble appears to be more successful with highly educated respondents, and when a probability wheel and color-coded cards are used. The time trade-off method, while expensive, has been found to be easier for respondents than the standard gamble [38].

In general, the category ratings and magnitude estimation methods are least expensive and easiest to understand. Little has been written about the feasibility of the equivalence method, other than Patrick *et al.*'s observation that it was too complex for use outside a laboratory [16]. Also, the unrealistic assumptions and emotive nature of the task confused and offended some judges. Because it is so similar conceptually to magnitude estimation one could speculate that many of the same strengths and weaknesses apply to both methods.

One indication of a scaling method's acceptability to respondents is response rate, although response rate is influenced by other variables as well. High response rates have been achieved with all methods. The willingness-to-pay method has suffered from low response rates (under 50%) in two studies [21, 47]. This has been explained on the basis that patients cannot understand the task, are hostile to the question, or have little idea of how much is spent on health care items [48]. However, in a recent study, Thompson [19] was able to achieve a 96% response rate, with 84% of respondents giving plausible answers. Both the likelihood of response and plausibility of response increased with education. Thompson attributes the high rate of response to improved questionnaire

design, improved performance of the interviewers, and having no subjects older than 66 years.

Response rate appears to be as much a function of population group as of scaling method. On the basis of his review of eight different studies, Torrance [12] reports that participation rates were lowest for the general public (70–80%), and highest for those with a special interest in research, like patients or clinicians (83–100%).

CONCLUSIONS

Based on data concerning their reliability, validity, and feasibility, the most promising scaling methods are the category ratings, magnitude estimation, and the time trade-off methods. The category ratings method is easiest to administer, and appears to yield scale values that are as valid as any other method. Thus, in large-sample studies, this would seem to be the method of choice.

Magnitude estimation is also relatively easy to administer. This method appears to yield valid scale values when 0 is defined as the absence of disease and disability, and the upper extreme is left open. This allows health states to be evaluated as worse than death. Magnitude estimation is recommended over category ratings in situations where a ratio-level scale is required, for example, when the investigator wants to be able to say that health-state A is twice as desirable as health-state B.

The time trade-off method is more expensive and difficult to administer than the other two methods, but several studies support its validity. Unlike the category ratings and magnitude estimation methods, it asks respondents to make a decision. Having to make a decision about the number of years one would give up to be in a healthy state may lead to more thoughtful consideration of each health state. However, a potential difficulty with the time trade-off is that individuals probably discount years in the future, viewing them as less important than current years. Thus, it cannot be assumed that every year "traded off" has the same value.

When the decision problem under study involves uncertainty, as do most clinical decisions, the standard gamble may have particular value due to its risk orientation, but it is not recommended for population studies because it is complex, expensive and difficult to administer. More research is needed to determine the

psychometric qualities of the equivalence and willingness-to-pay methods before they can be endorsed for use in health preference studies; however, both ask respondents to make choices they are often unable or unwilling to make. In particular, since the notion of equating a certain number of healthy people with a greater number of disabled persons is offensive to many respondents, we do not recommend the equivalence method.

Acknowledgements—The authors wish to express their appreciation to Allan Detsky, Walter Spitzer, and Eugene Rich for their helpful comments on an earlier version of this paper.

Editor's Note

This manuscript is the second of a four-part series, to be completed in the next two issues of the *Journal of Clinical Epidemiology*.

REFERENCES

1. McDowell I, Newell C. *Measuring Health: A Guide to Rating Scales and Questionnaires*. New York: Oxford University Press; 1987.
2. Carter WB, Bobbitt RA, Bergner M, Gibson BS. Validation of and interval scaling: the sickness impact profile. *Health Services Res* 1976; Winter: 516–528.
3. Nunnally JC. *Psychometric Theory*. New York: McGraw-Hill; 1978: 2nd edn.
4. Kaplan RM, Ernst JA. Do rating scales produce biased preference weights for a health index? *Med Care* 1983; XXI: 193–207.
5. Lipscomb J. Value preferences for health: meaning, measurement and use in program evaluation. In: Kane RL, Kane RA, Eds. *Values and Long Term Care*. Lexington, Mass.: Lexington Books; 1982.
6. Engen T. Psychophysics. In: Kling JW, Riggs LA, Eds. *Woodworth and Schlosberg's Experimental Psychology, Vol. 1: Sensation and Perception*. New York: Holt, Rinehart and Winston; 1972: 47–86.
7. Torgerson WS. *Theory and Methods of Scaling*. New York: John Wiley; 1958.
8. Torrance GW. Measurement of health state utilities for economic appraisal. A review. *J Health Econ* 1986; 5: 1–30.
9. Wendt D. On SS Stevens' psychophysics and the measurement of subjective probability and utility. In: Wegener B, Ed. *Social Attitudes and Psychophysical Measurement*. Hillsdale, N.J.: Lawrence Earlbaum; 1982.
10. von Neumann J, Morgenstern O. *Theory of Games and Economic Behavior*. New York: John Wiley; 1953.
11. Torrance GW, Thomas WH, Sackett DL. A Utility Maximization Model for evaluation of health care programs. *Health Service Res* 1972; 7: 118–133.
12. Torrance GW. Utility approach to measuring health-related quality of life. *J Chron Dis* 1987; 40: 593–600.
13. Stevens SS. Issues in psychophysical measurement. *Psychol Rev* 1971; 78: 426–450.
14. Kaplan RM, Bush JW, Berry CC. Health status index: category ratings versus magnitude estimation for measuring levels of well being. *Med Care* 1979; 17: 501–525.
15. Rosser R, Kind P. A scale of evaluations of states of illness: is there a social consensus? *Int J Epidemiol* 1978; 7: 347–358.

16. Patrick DL, Bush JW, Chen MM. Methods for measuring levels of well-being for a health status index. *Health Services Res* 1973; 8: 228-245.
17. Haig JHB, Scott D, Wickett LI. The rational zero point for an illness index with ratio properties. *Med Care* 1986; 24: 113-124.
18. Berg RL. Establishing the values of various conditions of life for a health status index. In: Berg RL, Ed. *Health Status Indexes*. Chicago: Hospital Research and Educational Trust; 1973.
19. Thompson MS. Willingness to pay and accept risks to cure chronic disease. *Am J Public Health* 1986; 76: 392-396.
20. Shepard DS, Zeckhauser RJ. Life-cycle consumption and willingness to pay for increased survival. In: Jones-Lee MW, Ed. *The Value of Life and Safety*. Amsterdam: North Holland; 1982: 95-141.
21. Thompson MS, Read JL, Liang M. Feasibility of Willingness-to-Pay measurement in chronic arthritis. *Med Decis Making* 1984; 4: 195-215.
22. Howard RA. The foundations of decision analysis. *IEEE Trans Systems Sci Cybern* 1968; SCC-4: 200-210.
23. North DW. A tutorial introduction to decision theory. *IEEE Trans Systems Sci Cybern* 1968; SCC-4: 200-210.
24. Luce RD, Raiffa H. *Games and Decisions*. New York: Wiley; 1957.
25. Veit T, Rose BJ, Ware Jr JE. Effects of physical and mental health on health state preferences. *Med Care* 1982; 20: 368-401.
26. Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF. The measurement of patients' values in medicine. *Med Decis Making* 1982; 2: 449-462.
27. Schoemaker PJH. The expected utility model: its variants, purposes, evidence and limitations. *J Econ Lit* 1982; XX: 529-563.
28. Kahneman D, Tversky A. The psychology of preference. *Sci Am* 1982; 246: 160-173.
29. Stevens SS. *Handbook of Experimental Psychology*. New York: John Wiley; 1951.
30. Stevens SS. To honor Fechner and repeal his law. *Science* 1961; 133: 80-86.
31. Stevens SS. A metric for the social consensus. *Science* 1966; 151: 530-541.
32. Anderson NH. How functional measurement can yield validated interval scales of mental qualities. *J Appl Psychol* 1976; 61: 677-692.
33. Weiss DJ. Quantifying private events: a functional measurement analysis of equisection. *Percept Psychophys* 1975; 17: 351-357.
34. Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC. Preferences for health outcomes: comparison of assessment methods. *Med Decis Making* 1984; 4: 315-329.
35. Christensen-Szalanski JJ. Discount functions and the measurement of patients' values: women's decisions during childbirth. *Med Decis Making* 1984; 4: 45-58.
36. O'Connor AM, Boyd NF, Warde P, Stolbach L, Till JE. Eliciting preferences for alternative drug therapies in oncology: influence of treatment outcome description, elicitation technique and treatment experience on preferences. *J Chron Dis* 1987; 40: 811-818.
37. Llewellyn-Thomas HA, Sutherland HJ, Ciampi A, Etezadi-Amoli J, Boyd NF, Till JE. The assessment of values in laryngeal cancer: reliability of measurement methods. *J Chron Dis* 1984; 37: 283-291.
38. Torrance GW. Social preferences for health states: an empirical evaluation of three measurement techniques. *Socioecon Planning Sci* 1976; 10: 129-136.
39. Wolfson AD, Sinclair AJ, Bombardier C, McGeer A. Preference measurements for functional status in stroke patients: interrater and intertechnique comparisons. In: Kane RL, Kane RA, Eds. *Values and Long Term Care*. Lexington, Mass.: Lexington Books; 1982.
40. Llewellyn-Thomas HA, Sutherland HJ, Tikshirani R, Ciampi A, Till JE, Boyd NF. Methodologic issues in obtaining values for health states. *Med Care* 1984; 22: 543-552.
41. Detsky A, McLaughlin JR, Abrams B, L'Abbe A, Whitwell J, Bombardier C, Jeejeebhoy KN. Quality of life of patients on long-term total parenteral nutrition at home. *J Gen Intern Med* 1986; 1: 26-33.
42. Churchill DN, Torrance GW, Taylor DW, Barnes CC, Ludwin D, Shimizer A, Smith EKM. Measurement of quality of life in end-stage renal disease: the Time Trade-Off approach. *Clin Invest Med* 1987; 10: 14-20.
43. Miles DL. Health care evaluation project terminal progress report. National Center for Health Services Res Grant 5 RO1 HS01568 1977; July.
44. Marks LE. Psychophysical measurement: procedures, tasks, scales. In: Wegener B, Ed. *Social Attitudes and Psychophysical Measurement*. Hillsdale, N.J.: Lawrence Earlbaum; 1982.
45. Kind P, Rosser R, Williams A. Valuation of quality of life: some psychometric evidence. In: Jones-Lee MW, Ed. *The Value of Life and Safety*, Amsterdam: North-Holland; 1982: 159-170.
46. Pauker SP, Pauker SG. The amniocentesis decision: Ten years of decision analytic experience. *Birth Defects* 1987; 23: 151-169.
47. Acton JP. Evaluating public programs to save lives: The case of heart attacks. The Rand Corporation, Santa Monica 1973; (Report R950RC).
48. Liang MH, Robb-Nicholson C. Health status and utility measurement viewed from the right brain: experience from the rheumatic diseases. *J Chron Dis* 1987; 40: 579-583.
49. Cadman D, Goldsmith C. Construction of social value or utility-based health indices: the usefulness of factorial experimental design plans. *J Chron Dis* 1986; 39: 643-651.
50. Churchill DN, Torrance GW, Taylor DW, Barnes CC, Ludwin D, Shimizer A, Smith EKM. Measurement of quality of life in end-stage renal disease: the Time Trade-Off approach. *Clin Invest Med* 1987; 10: 14-20.
51. O'Connor AM, Boyd NF, Till JE. Influence of elicitation technique, position order and test-retest error on preferences for alternative cancer drug therapy. *Nursing Research: Science for Quality Care; Proc 10th National Nursing Research Conference*. Toronto: University of Toronto; 1985.
52. Torrance GW. Preferences for health states: a review of measurement methods. *Mead Johnson Symposium on Perinatal and Developmental Medicine* 1982; 20: 37-45.

METHODOLOGY FOR MEASURING HEALTH-STATE PREFERENCES—III: POPULATION AND CONTEXT EFFECTS

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—In addition to the scaling method, there are many other aspects of the measurement process that may affect rater judgments of the relative desirability of health states. Although we find little compelling evidence of population differences in preferences due to demographic characteristics, there is some evidence suggesting that medical knowledge and/or experience with illness may influence raters' valuations of health states. Other aspects of the rating process that affect rater judgments can be classified as one of two types: inconsistencies due to limitations in human judgment, and inconsistencies due to situation-specific variables. When inconsistencies are due to limitations in human judgment, such as framing effects, a reasonable solution is to help the rater to see and correct the inconsistency. When inconsistencies are due to situation-specific variables, such as the way the health state is defined and presented, investigators should attempt to standardize conditions across studies.

Values preferences Preference weights Social preferences Utility measurement Health-state preferences Health status measurement

INTRODUCTION

In addition to the scaling method, there are many other aspects of the measurement process that may affect rater judgments of the relative desirability of health states. Evidence suggests that certain characteristics of the rater, such as medical knowledge or experience with an illness, may influence his or her judgments. Also, the way health states are defined, labeled, and presented has been demonstrated to influence rater judgments; even subtle changes in wording can produce preference reversals. In this section, we first review empirical findings on preference differences among population groups; then we discuss other context variables that affect rater judgments.

PREFERENCE DIFFERENCES AMONG POPULATION GROUPS

Several health status measures have made use of preference studies in order to assign values to multiattribute health states [1-3]. These preference-based health status measures are used to measure the outcomes of particular policies and programs. A number of questions can be raised concerning this application of preference weights, one of which is the appropriateness of aggregating preferences using the arithmetic mean, which we briefly introduced in Part I (J Clin Epidemiol 1989; 42: 345-354). Another question that has arisen in this context is whose valuations should be incorporated into an index. Some have argued that it may not matter whose preferences are used if it can be demonstrated that no major differences exist among groups of raters. Two general types of studies have been conducted to address this question: studies of variation across population

*Reprint requests should be addressed to: Debra Froberg, Ph.D., Division of Human Development and Nutrition, School of Public Health, University of Minnesota, Box 197 UMHC, 420 Delaware Street S.E., Minneapolis, MN 55455, U.S.A.

subgroups due to demographic characteristics, and studies of variation due to degree of medical knowledge or experience with an illness [4].

Demographic characteristics

Beginning with the first set of studies, we find little compelling evidence of population differences due to demographic characteristics. Numerous studies have found no differences in preferences attributable to sex or age [5-8]. The only exception is Sackett and Torrance's [9] finding that the utility values associated with 6 of their 15 disease-specific health states were associated with age. Older persons assigned lower utility to dialysis and transplantation, but higher utility to hospital confinement for an unnamed contagious disease.

Neither SES nor professional status appears to influence preferences [5, 7-9], nor do other demographic variables such as race, nationality, marital status, political persuasion, or religion [7, 8]. However, because some of the studies contain small numbers of subjects, and many showed a high degree of variability in the distribution of preferences, the results currently available may obscure meaningful differences among groups. Additional studies with adequate power to detect differences are needed to increase confidence that preferences do not depend upon demographic characteristics.

Medical knowledge/experience with illness

In contrast to the data on demographic characteristics, there is some evidence suggesting that medical knowledge and/or experience with illness may influence raters' valuations of health states. Sackett and Torrance [9] found that the health state of the respondent was related to utilities for some but not all health states; for example, home dialysis patients assigned higher utility to kidney dialysis than did the general public. This finding has prompted speculation that most patients with a particular disease or disability learn to cope with it, and therefore the general public's fear of and disutility for a condition may be exaggerated. In a more recent study, Llewellyn-Thomas *et al.* [10] reported that the rater's own health status did not influence ratings.

Carter *et al.* [6] compared the ratings of a group of health professionals (physicians, nurses and health administration students) with those of a random sample of enrollees of a prepaid health plan. Although the ordering of items did not differ, the consumer judges tended

to assign higher scale values than the health professionals. In a study of nursing home outcomes, Kane *et al.* [11] reported that the importance attributed to a particular health domain varied substantially with the type of respondent. In particular, significant differences were noted between nursing home residents, and non-residents; of the nonresident groups, family members' ratings deviated most from the overall mean ratings.

In two additional studies, some significant differences between respondent groups were found, but considering the total number of pairwise comparisons conducted, the number of significant differences was small. Among Wolfson *et al.*'s [4] 840 pairwise comparisons among physicians, physical and occupational therapists, family members of stroke patients, and stroke patients, only 15 pairs were statistically significant. If the significance level had been adjusted for the large number of comparisons, the number of significant findings would have been even fewer. Rosser and Kind [7] performed 14 pairwise comparisons among patients, nurses, physicians, and healthy volunteers and found two significant differences: medical patients vs physicians and medical patients vs psychiatric patients.

At this time, reports of no differences among rater groups outweigh those showing significant differences, although again, problems due to variability within groups and low statistical power may be obscuring differences. Preference patterns have been very similar among patients, physicians, and students [12, 13], between nursing students and visitors to a Cancer Institute open house [5], between students and health leaders [14], and between parents of chronically ill children and the general public [15]. Further, no differences were found among groups classified in terms of past experience as an inpatient, past experience of serious illness, history of severe pain, or family history of serious illness [7].

On the whole, the literature on rater differences suggests that while age and experience with the health state being rated (not general health status) may influence raters' valuations, the effects of most other demographic and experiential/medical variables are small or nonexistent. Even the evidence with respect to age and experience with health states is not overwhelming. We agree with Boyle and Torrance's [16] conclusion that "differences in valuations attributable to the personal characteristics of respondents are trivial when

compared with the differences that might arise from the alternative methodologies used to create an index in the first place" [16, p. 1054].

It should be emphasized that this does not mean people always express similar preferences for health states. In fact, Sackett and Torrance [9] reported a standard deviation of 0.30 for a distribution of health preferences on a 0–1 scale, indicating that respondents differed greatly in their preferences. Since empirical evidence suggests that these individual differences cannot be adequately explained by variables such as age, sex, socio-economic status, religion, illness, and other personal characteristics, the more important questions may involve the implications of using an average weight to represent a particular population. Perhaps, we should be as concerned about the variability of preferences within groups as we have been about variability between groups.

Returning to our original question, whose preferences should be measured? Since this is not an empirical question, research data can illuminate the issues but not provide a definitive answer. Fortunately, the bulk of the evidence points to no systematic preference differences among rater groups due to demographic characteristics. However, the finding that age and experience with the health state being rated *are* associated with preference values suggests that, in some cases, it may be appropriate to weight more heavily the preferences of those most directly affected by an intervention or policy. This seems especially true in clinical decision making, and may apply to some public policy decisions as well. However, there is considerable room for debate on this issue, as some believe that society's rather than patients' values should count when the general public is responsible for the cost.

It is clear that in addition to rater characteristics, many other aspects of the measurement process influence the quantitative results obtained, but what is less clear is whether these variations should be viewed as biases or as valid representations of the lability of value judgments. Although it is not always easy to distinguish between these two sources of inconsistency, we have tried to group studies on this basis. In the first group of studies, inconsistencies in preferences are viewed as errors in human judgment, whereas in the second group of studies, inconsistencies are attributed to the effects of valid independent variables, i.e.

situation-specific variables. We believe this distinction is helpful in determining whether and how to reconcile observed inconsistencies.

INCONSISTENCIES DUE TO LIMITATIONS IN HUMAN JUDGMENT

Most inconsistencies in preferences for health states that are due to limitations in human judgment arise when the same objective alternatives are viewed in relation to different points of reference. Tversky and Kahneman [17] have analyzed this phenomenon in a variety of situations, calling these inconsistencies "framing effects". For example, they show that when respondents are given a choice between two programs, they prefer one program when outcomes are defined in terms of the number of lives the program will save, but a different program when the same outcomes are defined in terms of the number of lives that will be lost. This reversal of preferences occurs despite the fact that the two situations are effectively identical. Certainly, preferences between options should not change with changes in frame, just as the perceived height of two neighboring mountains should not reverse with changes in vantage point. "Because of imperfections of human perception and decision, however, changes in perspective often reverse the relative apparent size of objects and the relative desirability of options" [17, p. 453].

If framing effects arise due to changes in reference point, what determines the rater's reference point? Sometimes it is the state to which one has adapted: the same tub of tepid water may be felt as hot to one hand and cold to the other if the hands have been exposed to water of different temperatures [18]. Often, however, reference points are provided to the rater by the investigator as in the example cited above in which outcomes were framed either in terms of lives lost or lives saved. In the particular context in which we are interested, namely the elicitation of preferences for alternative health states, the investigator may determine the reference point in at least three ways: (1) by providing anchors such as "perfect health" and "death", (2) by labeling diseases or treatments as opposed to leaving them unidentified, or, (3) by choosing a particular way of describing outcomes. A limited amount of empirical evidence exists suggesting that each of these ways of determining the rater's reference point does in fact influence preferences.

Anchoring effects

Sutherland *et al.* [19] found that values assigned to health states using rating scales were strongly influenced by the anchors on the scale. Compared to the values assigned to health states when the anchors consisted of perfect health and death, systematically higher values were assigned to the same states when the anchor of death was replaced by other states, and systematically lower values were assigned when the anchor of perfect health was replaced. Kaplan and Ernst [20] investigated context effects by comparing the ratings of different rater groups given only low, medium or high items and found little evidence of bias. Thus, while the scale anchors may influence ratings, the particular group of health states selected for rating does not appear to influence the ratings. In the case of magnitude estimation, the values obtained may be influenced by whether the "standard" health state comes from the middle or end of the scale [21].

Even the standard gamble has been shown to be internally inconsistent. In one study, the standard gamble yielded inconsistent results when other outcomes were substituted for the outcomes of perfect health and death [22]. According to expected utility theory, a rater's utility for a particular state should not be affected by changes in the gamble outcomes, just as a rater's values should not be influenced by the anchors in a rating scale. Hershey *et al.* [23] provide further evidence that variations in probabilities and outcome levels as well as other variations in the way the standard gamble is applied induce systematic bias in utility functions.

Labeling effects

The investigator may determine the rater's reference point through labeling as well as through anchoring. The way in which labeling can affect preferences is well documented in economic literature. Schoemaker [24] showed that a higher percentage of people preferred a sure loss of \$10 to a 1% chance of losing \$1000 when the scenario was labeled "insurance" than when it was labeled a "gamble". Similarly, consumers are more accepting of a cash discount than a credit card surcharge, even though the two labels differ only in terms of the implicit normal reference point [17]. In the health preference literature, two studies have indicated that labeling can make a difference in preference

values. In a study of clinical decision making, radiation therapy was chosen 42% of the time when it was not identified (referred to only as a treatment with specified outcomes) and only 26% of the time when it was identified [12]. Sackett and Torrance [9] found that labels had a significant effect on preferences; specifically, tuberculosis was preferred to an unnamed contagious disease and mastectomy for injury was preferred over mastectomy for breast cancer. However, one could argue that in both of these studies, labeling had the effect of providing more information to subjects about the health state; thus the resulting change in preferences should not be considered bias or error.

Outcome description effects

Several studies have shown that variations in the way outcomes are described can affect preferences. Twice it has been demonstrated that framing a clinical decision making problem in terms of the probability of dying produces different preferences than framing it in terms of probability of surviving [5, 12]. By using various combinations of positive, negative, and mixed frames, O'Connor *et al.* [5] concluded that the negative frame (probability of dying) appeared to be the biased one. In addition to the effects of the words dying and surviving, McNeil *et al.* [12] found that preferences were influenced by whether patients received cumulative probability data (probability of survival immediately after treatment and 5 years post), or life-expectancy data (probability of survival immediately after treatment and the life-expectancy associated with each treatment).

Other effects

Two additional variables that produce inconsistent preferences by changing the rater's reference point have been investigated. Llewellyn-Thomas *et al.* [10] found that mean scores assigned to narrative scenarios by category rating were substantially increased when the raters had first used the standard gamble. This effect was observed only with scenarios that were written in the first person singular, narrative form. There were no method sequence effects for the scenarios written in a standardized outline form. This finding is consistent with an earlier study in which no differences in preferences resulted from altering the order of presentation of the category rating and magnitude estimation methods for scenarios in standardized outline form [14]. Thus, on the basis of

this limited evidence, it appears that narrative-form scenarios are more susceptible to method sequence effects than are outline-form scenarios.

The effect of perceived prevalence of a disease on raters' judgment of its severity was examined by Jemmott *et al.* [25]. They found that subjects who thought the disease was more prevalent rated it as less serious than subjects who thought it was less prevalent. Whether this constitutes a bias is debatable, since there is some truth to the notion that serious diseases (especially fatal ones) are less prevalent than less serious diseases.

INCONSISTENCIES DUE TO SITUATION-SPECIFIC VARIABLES

Now we turn to aspects of the measurement process that we would *expect* to alter preferences, aspects that can be viewed as independent variables influencing a rater's true preferences. Three such variables are the prognosis and duration associated with health states, and the mode of presentation.

Prognosis and duration

Unfortunately, the field of health status measurement has been hampered by differences in the way investigators have handled prognosis and duration. Because of these differences, scale values for various multiattribute health indexes are not directly comparable. For example, scale values for the Sickness Impact Profile were obtained by asking judges to rate the severity of dysfunction described in an item without regard for what may be causing it. No mention is made of prognosis or duration [6]. On the other hand, Torrance *et al.* [3] asked subjects to imagine being in each state for a lifetime.

Kaplan *et al.* [21] argue that while knowledge of prognosis, the expected transitions across function levels over time, is essential to understanding the health status of an individual or group, prognosis should be separated from scale values of particular function levels in the measurement process. Thus, their Index of Well-Being is a static or time specific measure of function whereas the Weighted Life Expectancy incorporates the prognostic dimension.

Despite the lack of uniformity in the treatment of prognosis and duration by various investigators, only a few studies have been designed to identify the effects of these variables. In a study of scale values assigned to levels of disability and distress, Rosser and Kind

[7] found that changing the prognosis from treatable to permanent had very minor effects on scale values. In contrast, Sackett and Torrance [9] demonstrated that the utility assigned to a health state decreased as the duration of time in the state increased. Since these two studies were methodologically so different, particularly with respect to health-state descriptions and scaling methods, we cannot speculate about reasons for the contradictory findings. Further insight will require additional studies which systematically control selected variables.

Another variable that might be expected to influence rater preferences is whether raters evaluate the states in relation to themselves or to a hypothetical patient. In most studies, raters are either told or implicitly assume that the states apply to themselves. However, Ciampi *et al.* [26] investigated the effects on preferences of varying the characteristics of a hypothetical patient. A cancer patient was to be treated either conservatively without hope of cure, or radically by a risky treatment having cure or immediate death as possible outcomes. Variations in levels of the hypothetical patient's physical and psychosocial health and achievement motivation had a significant influence on preferences. Similarly, Kane *et al.* [11] found significant differences in the values respondents assigned to health outcomes depending upon whether the hypothetical nursing home patient was cognitively and functionally intact. The results of these studies highlight the serious ethical considerations that arise when social preferences are used to make public policy decisions.

Mode of presentation

Several studies have examined preference shifts due to the mode of presentation of health states. Preferences were not significantly influenced by the use of a computer compared with paper and pencil techniques [5]. However, differences were noted in two separate investigations when the mode of presentation resulted in different information being presented to raters. Boyd *et al.* [27] compared the preference values assigned to health states for (1) scenarios relating to laryngeal cancer patients' ability to carry out various activities and (2) a combination of the scenario and a voice recording. They found that scores assigned to the scenarios alone differed significantly from those assigned to the combination. In some cases scenarios alone were rated higher than the combined

scenario/voice recording, whereas in other cases the reverse was true.

In another study [10], two types of scenarios were used: a standardized outline form describing patients according to age, mobility, physical and social activity, and predominant symptom and/or problem; and a narrative form written in the first person singular. The information contained in the narrative form was more specific than in the outline form, and it also included more problems. Not surprisingly, the narrative form consistently received lower mean scores. Because the information presented in the alternative formats used in these two studies was substantially different, it is not possible to isolate the format effect. To do so would require that everything except the format remain essentially the same. For this reason, we include these studies among the group reflecting independent variables influencing true preferences rather than viewing these inconsistencies as errors in judgment.

In neither of these studies were preference values produced by different formats compared to a criterion; indeed, a criterion for health-state preferences has proven difficult to find. This leaves us in the dilemma of not knowing which type of format produces the most valid preference values. In the absence of such information, we surmise that moderately detailed health-state descriptions yield more accurate judgments of preference than either very scant descriptions or very lengthy descriptions that run the risk of overloading the rater's information processing capacity.

WHAT TO DO ABOUT CONTEXT EFFECTS

The distinction between inconsistencies due to errors in human judgment and those due to valid situation-specific variables is useful as we consider ways of reconciling inconsistencies. There is general agreement in the literature that when inconsistencies are due to human error, such as when the framing of a decision problem influences the rater's reference point, a reasonable solution is to help the rater to see and correct the inconsistency. Tversky and Kahneman [17] summarize the situation as follows:

Individuals who face a decision problem and have a definite preference (i) might have a different preference in a different framing of the same problem, (ii) are normally unaware of alternative frames and of their potential effects on the relative attractiveness of

options, (iii) would wish their preferences to be independent of frame, but (iv) are often uncertain how to resolve detected inconsistencies [17, pp. 457-458].

Thus, a strategy for eliciting consistent preferences is to seek convergent validation of preferences by presenting the problem in more than one way and asking the rater to reconcile any incompatibilities [3, 12, 23, 28]. For example, outcomes can be described in terms of both lives lost and saved, both 5-year survival rate and life expectancy, and both probability of surviving and probability of dying. Several investigators recommend that interviewers assume an active role in helping raters to clarify and correct incompatible responses. Thompson [28] reported that three interviewer interventions—providing explanatory introductions, repeating questions for initially baffled subjects, and allowing subjects to revise earlier answers—dramatically increased the number of plausible responses to the willingness-to-pay technique. However, when an interviewer takes an active role, the potential for influencing the rater is increased; care must be taken to minimize this bias. This can be done by using well-structured interviews and allowing clarification and elaboration only within narrow limits. Standard guidelines for training interviewers should be followed, such as role playing, conducting practice interviews, and assessing inter- and intrarater reliability.

Investigators should be aware of anchoring effects and deliberately select anchors that are appropriate to their application. If an investigator intends to compare his or her results to those of previous studies, the anchors must be the same. Since studies have shown that subjects will rate some states worse than death when given the opportunity to do so, future studies should allow for this possibility.

When inconsistencies are due to valid situation-specific variables, the objective is no longer to reconcile inconsistencies. Rather, it is to understand the relevant variables through conducting research and developing an explanatory theory. A major barrier to understanding situation-specific variables is our present lack of theory. A great deal of work has been done in testing expected utility theory, the prominent theory of decision making under uncertainty, and modifications in the theory have been proposed that recognize context effects. (See, for example, Kahneman and Tversky's [29] descriptions of prospect theory.) However, the

expected utility model does not adequately describe problem representation and will therefore not easily predict new context effects. Further, evidence that people make decisions contrary to the predictions of expected utility theory is so strong that some have argued that it has little relevance as a descriptive theory despite its usefulness as a prescriptive theory. To predict new context effects we need to better understand the psychological processes inherent in decision making [24].

One issue of critical importance to the measurement of health preferences is, what happens when people do not know, or have difficulty appraising what they prefer? Under these circumstances, elicitation procedures may become major forces in shaping the preferences expressed [30]. In addition, how do preferences elicited under research conditions compare with those expressed in emotionally-charged real-life situations? How do preferences obtained under conditions of social isolation, as in research settings, compare with those obtained after consultation with relatives, friends, and health professionals?

Research that clarifies influences on preferences and the decision making process in general will contribute much to advancing the field of health preference measurement. In the meantime we can proceed on the basis of our present knowledge of context effects. When inconsistencies result from judgment errors, interviewers can help raters to resolve them. When inconsistencies result from the effects of situation-specific variables, we can attempt to standardize conditions across studies, or if that is not desirable or feasible, we should view preferences as having validity only within the context in which they were measured.

Acknowledgements—The authors wish to express their appreciation to Allan Detsky, Walter Spitzer, and Judith Garrard for their helpful comments on an earlier version of this paper.

Editors' Note

This manuscript is the third of a four-part series, to be completed in the next issue of the *Journal of Clinical Epidemiology*.

REFERENCES

- Bergner M, Bobbitt RA, Carter W, Gilson BS. The sickness impact profile: development and final revision of a health status measure. *Med Care* 1981; XIX: 787-805.
- Patrick DL, Bush JW, Chen MM. Toward an operational definition of health. *J Health Soc Behav* 1973a; 14: 6-23.
- Torrance GW, Boyle MH, Horwood SP. Application of multi-attribute utility theory to measure social preferences for health states. *Oper Res* 1982; 30: 1043-1069.
- Wolfson AD, Sinclair AJ, Bombardier C, McGeer A. Preference measurements for functional status in stroke patients: interrater and intertechnique comparisons. In: Kane RL, Kane RA, Eds. *Values and Long Term Care*. Lexington, Mass.: Lexington Books; 1982.
- O'Connor AM, Boyd NF, Till JE. Influence of elicitation technique, position order and test-retest error on preferences for alternative cancer drug therapy. *Nursing Research: Science for Quality Care, Proc 10th National Nursing Research Conference*. Toronto: University of Toronto; 1985.
- Carter WB, Bobbitt RA, Bergner M, Gibson BS. Validation of and interval scaling: the sickness impact profile. *Health Serv Res* 1976; Winter: 516-528.
- Rosser R, Kind P. A scale of evaluations of states of illness: is there a social consensus? *Int J Epidemiol* 1978; 7: 347-358.
- Kaplan RM, Bush JW, Berry CC. The reliability, stability, and generalizability of a health status index. *Proc Social Statistics Section*. American Statistical Association; 1978: 704-709.
- Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. *J Chron Dis* 1978; 7: 347-358.
- Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF. Methodologic issues in obtaining values for health states. *Med Care* 1984; 22: 543-552.
- Kane RL, Bell RM, Reigler SZ. Value preferences for nursing home outcomes. *Gerontologist* 1986; 26: 303-308.
- McNeil BJ, Pauker SG, Sox HC, Tversky A. On the elicitation of preferences for alternative therapies. *N Engl J Med* 1982; 306: 1259-1262.
- Boyd NF, Sutherland HJ, Ciampi A, Tibshirani R, Till JE, Harwood A. A comparison of methods of assessing voice quality in laryngeal cancer. In: *Choices in Health Care: Decision Making and Evaluation of Effectiveness*. Toronto: Department of Health Administration, University of Toronto; 1982: 141-144.
- Patrick DL, Bush JW, Chen MM. Methods of measuring levels of well-being for a health status index. *Health Serv Res* 1973b; 8: 228-245.
- Cadman D, Goldsmith C. Construction of social value or utility-based health indices: the usefulness of factorial experimental design plans. *J Chron Dis* 1986; 39: 643-651.
- Boyle MH, Torrance GW. Developing multiattribute health indexes. *Med Care* 1984; 22: 1045-1057.
- Tversky A, Kahneman D. The framing of decisions and the psychology of choice. *Science* 1981; 211: 453-458.
- Kahneman D, Tversky A. The psychology of preference. *Sci Am* 1982; 246: 160-173.
- Sutherland HJ, Dunn V, Boyd NF. Measurement of values for states of health with linear analogue scales. *Med Decis Making* 1983; 3: 477-487.
- Kaplan RM, Ernst JA. Do rating scales produce biased preference weights for a health index? *Med Care* 1983; XXI: 193-207.
- Kaplan RM, Bush JW, Berry CC. Health status index: category ratings versus magnitude estimation for measuring levels of well being. *Med Care* 1979; 17: 501-525.
- Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF. The measurement of patients' values in medicine. *Med Decis Making* 1982; 2: 449-462.

23. Hershey JC, Kunreuther HC, Schoemaker PJH. Sources of bias in assessment procedures for utility functions. *Management Sci* 1982; 28: 936-954.
24. Schoemaker PJH. The expected utility model: its variants, purposes, evidence and limitations. *J Econ Lit* 1982; XX: 529-563.
25. Jemmott JB III, Ditto PH, Croyle RT. Judging health status: effects of perceived prevalence and personal relevance. *J Pers* 1986; 50: 899-905.
26. Ciampi A, Silberfeld M, Till JE. Measurement of individual preferences. *Med Decis Making* 1982; 2: 483-495.
27. Boyd NF, Sutherland HJ, Ciampi A, Tibshirani R, Till JE, Harwood A. A comparison of methods of assessing voice quality in laryngeal cancer. In: **Choices in Health Care: Decision Making and Evaluation of Effectiveness**. Toronto: Department of Health Administration, University of Toronto; 1982: 141-144.
28. Thompson MS. Willingness to pay and accept risks to cure chronic disease. *Am J Public Health* 1986; 76: 392-396.
29. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrika* 1979; 47: 263-291.
30. Fischhoff B, Slovic P, Lichtenstein S. Knowing what you want: measuring labile values. In: Wallsten TS, Ed. **Cognitive Processes in Choice and Decision Behavior**. Hillsdale, NJ: Erlbaum; 1980.

Practical Pharmacoeconomics

Determining unit cost values for health care resources in pharmacoeconomic studies

Daniel C. Malone, PhD, Sean D. Sullivan, PhD, and David L. Veenstra, PharmD, PhD

ABSTRACT

The purpose of pharmacoeconomics is to enhance pharmaceutical decision-making within a patient population. One major and ongoing criticism decision-makers have about pharmacoeconomic studies is that the costs used in the studies don't apply to "their" environment. Obtaining cost estimates for many types of commonly used health care resources can be difficult. This article reviews common and not-so-common sources of cost information within three main health care resource areas often involved in pharmacoeconomic evaluations: pharmaceuticals, physician services, and hospitalization.

(*Formulary* 2001;36:294-304.)

This column alternates with the Practical Issues in Outcomes Management column. Practical Pharmacoeconomics department editor **Edward P. Armstrong, PharmD**, is associate professor, department of pharmacy practice, college of pharmacy, the University of Arizona, Tucson. Department assistant editors are **Amy J. Grizzle, PharmD**, assistant director, Center for Health Outcomes and Pharmacoeconomic Research, the University of Arizona and **Daniel C. Malone, PhD**, assistant professor, the University of Arizona.

The authors of this column are **Dr. Malone, Dr. Sullivan**, associate professor, schools of pharmacy and public health and community medicine, University of Washington, Seattle; and **Dr. Veenstra**, assistant professor, school of pharmacy, University of Washington.

The primary purpose of pharmacoeconomics is to enhance pharmaceutical decision-making within a patient population. An ongoing criticism of pharmacoeconomic studies is that the study costs used don't apply to the end-users' environment. Pricing health care resources is one of the more difficult aspects of conducting a pharmacoeconomic study. Most researchers agonize over what costs are available and how the results can be generalized or targeted to cover the myriad of health care organizations and contracts.

In the United States, there is no "one" source for obtaining health care costs. In contrast, standardized costs are often available in countries with single-payer systems. For example, the Australian Pharmacy Benefits Advisory Council requires pharmaceutical manufacturers to submit a formulary application that includes standardized costs for various types of health care, including hospital inpatient care, hospital outpatient services, physician visits, and medications.¹

This article discusses the various methods of assigning a cost for three main areas of health care resources—pharmaceuticals, physician services, and hospital costs. Various cost sources are discussed along with where an investigator may find data, often free of charge.

Pharmaceutical Costs

As is commonly known, there is no one price for pharmaceuticals that can be applied to all situations. In the United States, pharmaceutical prices are determined by supply, demand, contracts, and rebates. The average wholesale price (AWP) is a frequently used cost in many pharmacoeconomic studies; most likely because it is readily available.^{2,3}

Sources of AWP. One source of AWP is *Drug Topics Red Book* (see table 1). The *Red Book* is a compendium of FDA-approved products and prices published by Medical Economics.² It also contains prices for those manufacturers who sell directly to pharmacies. Another source for AWP is First DataBank's master drug data base (MDDDB) or national drug data file (NDDF). These databases contain pricing information, including wholesale acquisition cost, as well as National Drug Codes (NDC) and maximum allowable cost schedules used by the Health Care Financing Administration (HCFA) for Medicaid. A major drawback to using either the MDDDB or NDDF is the price, which can range from \$3,000 to \$18,000 per year depending on the particular needs of the analyst. First DataBank also publishes a paper version, called *Price Alert*, which costs substantially less (\$129/year).

Another source of AWP information (until recently) was Multum, a Denver-based drug information company. The Multum relational database was available for downloading free of charge after completing an online agreement. Included in this database was NDC, drug name, generic drug name, and wholesale price. As of February 1, 2001, Multum removed AWP from the downloadable files, but nonprofit entities may still obtain this information by contacting Multum. NDC files are available free of charge.

Other sources of drug prices. An alternative to AWP is the wholesale ac-

TABLE 1

WHERE TO FIND PHARMACEUTICAL COST INFORMATION

PHARMACEUTICAL COST INFORMATION

Source	Where available	Type of information provided	Comments
<i>Drug Topics Red Book</i>	In print from <i>Medical Economics</i>	AWPs, direct prices to pharmacies	Useful when practice-specific costs are unavailable. Annual subscription is about \$200.
Master drug database National drug data file (First DataBank)	www.firstdatabank.com	AWPs, wholesale acquisition prices, NDCs*, MAC* schedules	Useful when practice-specific costs are unavailable. Annual subscription varies by data elements and use.
Multum	www.multum.com	NDCs, wholesale prices†	NDC files can be downloaded and imported into database programs such as Microsoft Access. AWP information is no longer provided in the dataset.
Drugstore.com Walgreens.com	www.drugstore.com www.Walgreens.com	retail drug prices retail drug prices	Free source of retail price. Cost includes dispensing fee and can be less than prevailing local market prices.
Pharmacy Benefits Management Strategic Health Group	www.vapbm.org	Federal Supply Schedule for medications, NDCs	Useful for obtaining minimum drug acquisition prices. Database files can be downloaded free of charge.
NDC Health Information Services IMS America	www.simatics.com us.imshealth.com	for both: drug market share, average retail price and quantity	Useful for weighting analyses by drug market share.

* NDC = national drug codes; MAC = maximum allowable cost for Medicaid
† AWP is still available to nonprofit entities

Formulary/Source: D.C. Malone, PhD, S.D. Sullivan, PhD, and D.L. Veenstra, PharmD, PhD

quisition cost, which is more reflective of prices paid by pharmacy wholesalers. Wholesale acquisition costs can be found in either the MDDB or NDDF files, but these costs represent *estimated* acquisition costs, actual acquisition costs will vary by wholesaler.

The retail sector offers another source of drug pricing information. Retail prices include both the acquisition cost and the cost of dispensing, assuming the pharmacy is not using a particular product as a loss leader. With the advent of Internet-based retail pharmacy, it is now relatively easy to obtain a "market" price by logging onto various internet pharmacy sites, such as

Drugstore.com or Walgreens.com.

The federal government provides yet another source of drug cost information. Under the Omnibus Budget Reconciliation Act of 1990, the federal government required pharmaceutical manufacturers to provide Medicaid agencies the "best price" for pharmaceuticals or 15.1% off of the average manufacturer price to wholesale distributors.^{4,5} For the Department of Veteran Affairs, the Department of Defense, the Public Health Service, the Coast Guard, and the Indian Health Service, the price paid for pharmaceuticals is published in the federal supply schedule (FSS). The FSS represents ei-

ther "the same discount off of a drug's list price that the manufacturer offers its most-favored nonfederal customer under comparable terms and conditions" or 24% off their nonfederal average manufacturer price.⁵ The Pharmacy Benefits Management Strategic Health Group at the Hines Veterans Affairs Medical Center in Chicago is responsible for maintaining the government's federal supply schedule (FSS) for medications.

The FSS is available as a downloadable database file containing approximately 16,000 products for which the government has a contract. The file can be loaded into a database program, such

as Microsoft Access, and individual drug products can be searched by brand or generic name or NDC.

Issues about use of AWP and other confounding variables. Of course everyone knows that "wholesale" price is a misnomer. Pharmaceutical purchasers typically pay between 10% to 17% less than wholesale cost. Thus, using AWP is not optimal because it does not represent the true transaction cost. The price paid by pharmacies for a product, for example, doesn't reflect the cost to prepare and dispense a medication, which has been estimated to be between \$5.17 to \$6.77.^{6,7} Costs for medications administered in hospitals and physician offices are likely to be bundled with other services and therefore are much harder to estimate. Using a discounted price off AWP is an alternative, but raises another set of issues.

Another complicating factor is the variance of pricing by package size. In most cases, larger package sizes are less expensive per unit than smaller package sizes. Generic drugs and multiple source products create another dilemma, as the price may vary substantially across the various manufacturers and package sizes.

Drug cost information: Considerations for conducting pharmacoeconomic studies. There is no single best source for determining drug cost information for use in all pharmacoeconomic studies. To illustrate how costs vary per source: 10 mg atorvastatin, packaged with 90 units per package, is \$1.15 per tablet according to the FSS. The AWP and wholesale price, from Multum, is \$1.88, and \$1.67, respectively. The decision of which source to use is largely determined by practice setting. It might be most relevant for hospitals and managed care organizations to use their own drug acquisition costs. Researchers at PBMs and others who don't have access to health care organizations' drug acquisition costs, those conducting multicenter studies, or those who wish to conduct studies from a best price or governmental perspective could start with the FSS for medications. Using this database as a baseline, one could estimate the mini-

The complexity of pharmacoeconomics: An example

Determining health care resource use in and of themselves is a difficult task. Further compounding the difficulty is identifying and applying the appropriate costs to meet the specific situation. To illustrate this added level of complexity, let's look at the important distinction between *average cost* and *marginal cost*. As an example, the average cost of a day in the ICU could be calculated by summing all expenditures associated with the ICU over a defined period and then dividing the result by the total number of patient care days. To examine the cost effectiveness of the new drug that reduces GI bleeding, a researcher might look at length of ICU stay as one possible end point.

The problem with using the average cost of an ICU day in the above scenario is that it assumes that resource consumption remains the same across all ICU days. This assumption is probably not accurate because a patient admitted with a bleeding ulcer will consume more resources initially when they are undergoing diagnostic evaluation and early treatment. As the hemorrhaging is controlled, fewer resources would be used.

If cost per day was graphed, the line would be downward sloping be-

cause the cost per day is diminishing (marginal cost curves are actually J-shaped, but assume that hospitals have an incentive, such as a capitated payment, to discharge the patient as soon as possible.)

Under the assumption that the first day of care costs \$5,000 and the last day of care costs only \$1,000, it is important to make a decision regarding what portion of the ICU stay will be avoided when a new medication is administered. Assuming the medication saves on the more intensive portion of the stay, the resulting offset would be \$5,000 per day, compared with only \$1,000 per day if the drug affects the last day of care. Thus, the marginal cost is the more appropriate cost to include in the pharmacoeconomic analysis.

That said, it is rare to find studies that apply unit costs at the margin. Given the above example, one could imagine the detailed costing that must be performed to arrive at an estimate of the marginal cost. To compound the problem, many health care organizations have not conducted detailed cost-accounting studies to permit such a costing technique.

num acquisition price for various pharmaceuticals and then apply the appropriate adjustment based on practice setting or study perspective.

Deciding which source of cost information to use is only the first step in determining the price. If the drug to be studied is available only from a single source (ie, brand-name product), then it might make sense to use the lowest cost per unit in the analysis (since costs vary by package size). If the drug to be studied is available from multiple sources, the simplest solution might be to take an average of all manufacturers, using the lowest cost per unit for each manufacturer, regardless of package size. This assumes the researcher

has access to a database like the MDDB or to the *Red Book*, which contains a listing for each manufacturer and every package size.

Even with access to such a database, this method ignores the various market shares among the manufacturers. To obtain a cost based upon market share—particularly when conducting modeling studies—one should turn to the NDC Health Information Services or IMS America. These firms track prescription drug market sales and provide data that would allow a researcher to calculate a weighted cost based upon the product, average units per prescription, and number of prescriptions dispensed.

For example, Gonzales et al⁸ con-

TABLE 2

WHERE TO FIND PHYSICIAN AND HOSPITAL COST INFORMATION

Source	Where available	Type of information provided	Comments
Physician costs			
HCFA	www.hcfa.gov/stats/pufiles.htm relative value scale	Medicare reimbursement rates based upon resource-based upon type of service provided. Files are available for free.	Establishes reimbursement rates to physicians based
Hospital costs			
HCFA	www.hcfa.gov/stats/pufiles.htm	diagnosis-related group relative weight file downloadable free of charge.	Medicare's payment schedule based upon DRG. File is
American Hospital Directory	www.ahd.com care hospitals	Utilization and cost for Medicare patients at specific acute more detailed information.	Some data provided free of charge; fee charged for

Formulary/Source: D.C. Malone, PhD, S.D. Sullivan, PhD, and D.L. Veenstra, PharmD, PhD

ducted a study using the 1997 National Ambulatory Medical Care survey (NAMCS), which asked participating physicians to document a series of patient encounters. The name of each prescription medication was included in the documentation for each encounter. Unfortunately, no information was available on prescription strength or quantity. To estimate the cost of each medication, market share and average price per prescription was obtained from the NDC Health Information Services Source Prescription Database. Prescription cost estimates were then weighted taking into account the complex sampling design employed in NAMCS.

Physician Costs

Another component that may have to be accounted for in some pharmacoeconomic studies is physician costs. Although well known by some readers, for a complete discussion, a brief review of billing procedures follows.

Physicians typically bill for their services using the American Medical Association's current procedural terminology (CPT) codes.⁹ Insurance claims for physician services include the diagnosis in International Classification of

Diseases, 9th Revision, Clinical Modification (ICD 9 CM) format. Thus, given a claims file of physician services, a researcher could select the condition of interest using ICD 9 CM codes and the corresponding billed or reimbursed amount. In the absence of claims data, one can estimate the cost of physician services using HCFA fee schedules linked to CPT codes.

CPT codes are grouped into six broad categories: evaluation and management (99201 to 99499); anesthesiology (00100 to 01999, 99100 to 99140); surgery (10040 to 69979); radiology (70010 to 79999); pathology and laboratory (80002 to 89399); and medicine, excluding anesthesiology (90701 to 99199).⁹ In addition to the five digit code, a two digit modifier can be added to show that a procedure has been altered in some manner. Modifiers can be included when a procedure was performed by more than one physician, when a procedure was increased or reduced, when only part of a procedure was performed, or when an unusual event or complication occurred.

CPT codes are designed to reflect the workload associated with each patient encounter. Evaluation and management codes account for obtaining relevant pa-

tient history, physical examinations, medical decision making, counseling, coordination of care, nature of the presenting problem, and time spent addressing the problem. For example, the CPT code 90213 is for an established patient seen at the office or other outpatient facility and requires two of the three following components: an expanded problem focused history, an expanded problem focused examination, or medical decision making of low complexity. An example of a patient encounter with this code is: "Office visit with 55-year-old male, established patient, for management of hypertension, mild fatigue, on beta blocker/thiazide regimen."

A CPT code of 99214 is also for an office or outpatient visit and requires two of three of the following: a detailed history, a detailed examination, or medical decision making of moderate complexity. An example for this code is: "Office visit with 50-year-old female, established patient, diabetic, blood sugar controlled by diet. She now complains of frequent urination and weight loss, has a blood sugar of 320 and negative ketones on dipstick."

Specific procedures conducted in the physician office are linked to CPT

codes. For example, a visit for an asthma patient that is assessed for oxygen saturation via ear or pulse oximetry would have a CPT code of 94760.⁹

In an effort to stem the rising cost of physician services and also to pay physicians of various specialties similar amounts for providing the same service, HCFA developed resource-based relative value unit (RBRVU) for physician payments under Medicare.¹⁰ Based upon CPT codes, the RBRVU takes into account physician work, malpractice, and office expense to derive a weight associated with each CPT code. The RBRVU weights are published in the Federal Register and on the HCFA Web site (see table 2 for Web address). A group of files, including a Microsoft Excel spreadsheet and supporting documentation, can be downloaded free of charge (RVU01. EXE for year 2001). Each user must agree to abide by the terms and conditions associated with using the AMA's copyrighted CPT codes.

To calculate a physician's payment, each component (work, office expense, and malpractice) of the RBRVU is weighted by a geographical adjustment and then multiplied by the conversion factor. For 2001, the conversion factor is \$38,2581. Assuming the geographical weights are 1 for each component, the amount paid by Medicare for CPT code 99214 would be \$78.81 (2.06 [the RBRVU relative weight] × \$38,2581).

A single office visit, of course, may be associated with more than one CPT code. For example, a Medicare patient with asthma who presents to a physician's office (99214, CPT code for office visit) with an acute respiratory exacerbation may be assessed via pulse oximetry for oxygen saturation (94760, CPT code for pulse oximetry) and receive albuterol via nebulization (94640, CPT code for nonpressurized inhalation treatment for acute airway obstruction) in addition to other evaluation and management. The total cost of this episode of care would be [(2.06 × \$38,2581) + (0.16 × \$38,2581) + (0.69 × \$38,2581)] = \$111.33.

For comparison purposes, physician claims data from a large western United States health maintenance organization

for these same CPT codes were evaluated. The average payment using this combination of CPT codes was \$111.72, reflecting that Medicare fee schedules can be very similar to prevailing market rates.

Hospital Costs

Starting in 1983, Medicare began paying hospitals based on diagnosis-related groups (DRG). The DRG relative weight file can be downloaded from the HCFA Web site (see table 2).

In addition to each DRG having a weight, each facility also has a weight

Having a greater understanding of cost sources will better equip researchers and decision-makers to interpret and apply the results of pharmacoeconomic studies.

to adjust for various hospital characteristics including disproportionate share, urban or rural status, geographic region, wage index, indirect medical education, and several other factors.

Once the appropriate DRG weight is determined, it can be multiplied by a conversion factor to obtain the reimbursed amount. The conversion factor is updated annually and can be obtained from Medicare fiscal intermediaries for the various geographical regions of the United States. Information on the Medicare fiscal intermediaries can be found on the HCFA Web site.

For acute care facilities in the United States, the American Hospital Directory uses Medicare data to provide information on specific institutions (<http://www.ahd.com>). The Web site provides a free service that allows interested parties to examine utilization and costs associated with specific hospitals. Each report based upon HCFA data displays major service units (eg, cardiology, gynecology, medicine), the

number of Medicare inpatients for fiscal year 1999, average length of stay, average charges, and the Medicare case mix index. More detailed information, including cost-to-charge ratios can be purchased through the American Hospital Directory.

Miscellaneous HCFA Files

The HCFA offers other downloadable data sets (see table 3) that might be of interest to researchers conducting pharmacoeconomic studies. For example, the laboratory fee schedule paid by Medicare can be downloaded. Laboratory costs are linked to laboratory-CPT (LCPT) codes. Similar to physician CPT codes, LCPT codes are owned by the American Medical Association and users must agree to terms and conditions of use posted on the HCFA Web site prior to downloading files. In addition to a national payment, the fee schedule also contains median payment limits for each state.

Other files available from HCFA include those pertaining to home health agencies, skilled nursing facilities, ambulance services, and a durable medical equipment (DME) schedule. Prices for home health services reimbursed by HCFA are available in the public use files section of the HCFA Web site. Services provided by various providers (eg, nursing, physical therapy, occupational therapy) are separated out. Some durable medical equipment is also covered for Medicare recipients and reimbursed costs are provided in the payment to noninstitutional providers section. Examples of products covered include blood glucose reagent strips, tape, gauze, nebulizer, and IV pole. The file contains ceiling and floor prices as well as state scheduled amounts. Medicare reimburses at 80% of either the actual charge or the ceiling price, whichever is lower. The DME fee schedule is updated quarterly.

Conclusion

The lack of a uniform payer for health care in the United States creates a serious dilemma for researchers conducting and interpreting pharmacoeconomic studies. There is no one best method to price various health care resources.

TABLE 3
OTHER HCFA FILES AVAILABLE

Source	Where available	Type of information provided	Comments
Home health care			
HCFA	www.hcfa.gov/stats/pufiles.htm	Medicare average costs for home health services by facility.	Contains cost per skilled nursing visit, physical therapy visit, occupational therapy, speech pathology, and medical social services.
Skilled nursing facilities (SNF)			
HCFA	www.hcfa.gov/stats/pufiles.htm	Medicare reimbursement rates	Contains cost data for a variety of services administered in SNF facilities.
Durable medical equipment			
HCFA	www.hcfa.gov/stats/pufiles.htm	Medicare reimbursement rates	Contains durable medical equipment, prosthetics/orthotics, and supplies fee schedule.
Laboratory and diagnostics			
HCFA	www.hcfa.gov/stats/pufiles.htm	Medicare reimbursement rates	Contains reimbursement rates for various laboratory and diagnostic tests linked to L-CPT codes.

Formulary/Source: D.C. Malone, PhD, S.D. Sullivan, PhD, and D.L. Veenstra, PharmD, PhD

Organizations such as the International Society for Pharmaceutical Outcomes Research may be an appropriate venue for developing a standardized cost schedule. In lieu of this, a recommended method is to use costs from the organization from which the effectiveness or efficacy data are derived.

Organizational factors may often affect the cost and use of health care products and/or services. In the absence of organizational cost data, we recommend using the Medicare fee schedules. These schedules are publicly available and are transparent to all end-users, thus avoiding the issue of proprietary data concerns. For determining the price of medications, AWP minus 15% plus a dispensing fee is typically more appropriate than using only AWP for medications taken primarily by ambulatory populations. However, it might be prudent to use the FSS, which is ef-

fectively the best price for any purchaser. If these medication costs were used and the product being evaluated was considered to be cost effective, the results would be more convincing.

Despite the challenges discussed in this article, having a greater understanding of the cost sources available in the United States will better equip researchers and decision-makers to interpret and apply results from pharmacoeconomic studies.

REFERENCES

1. Commonwealth Department of Health HL-GaCS. Manual of Resource Items and their Associated Costs. 1993. Canberra, Australia, Australian Government Publishing Service.
2. RedBook. Montvale, NJ: Medical Economics, 2000.
3. www.firstdatabank.com, accessed 2-12-01.
4. Anonymous. Medicaid best price provisions fall short, official testifies at hearing on Medicare drug benefit. *Am J Health Syst Pharm* 2000;57:1028.
5. Danzon PM. Price comparisons for pharmaceuticals: A review of US and cross-national studies. Washington DC: The AEI Press, 1999.
6. Huey C, Jackson RA, Pirl MA. Analysis of the impact of third-party prescription programs on community pharmacy. *J Res Pharmaceutical Economics* 1995;6:57-72.
7. Schafermyer KW, Schondelmeyer SW, Thomas J, Proctor KA. Analysis of the cost of dispensing third-party prescriptions in chain pharmacies. *J Res Pharmaceutical Economics* 1992;4:3-24.
8. Gonzales R, Malone DC, Maselli JH, Sande MA. Excess antibiotic use for acute respiratory infections in the United States. *Clinical Infectious Disease* (in press).
9. American Medical Association, CPT 2001 Professional Edition. Chicago, IL: American Medical Association, 2001.
10. Hsiao WC, Braun P, Dunn DL, et al. An overview of the development and refinement of the Resource-Based Relative Value Scale. The foundation for reform of U.S. physician payment. *Medical Care* 1992;30 (11 Suppl):NS1-12. #

3. Health Utilities Index (Mark III)

The Health Utilities Index has been modified recently to include the following eight attributes. No classification for this version is available yet, but the attributes are:

1. Vision
2. Hearing
3. Speech
4. Getting around (mobility)
5. Hands and fingers (dexterity)
6. Feelings (emotional function)
7. Memory and thinking (cognitive function)
8. Pain and discomfort

A questionnaire from the Ontario Health Survey that can be used to collect data for this Index follows. Utility weights for this eight-attribute index are currently being collected. For further information about the utility weights and calculating formulas available in the future, please contact Dr. George Torrance.

Ontario Health Survey Interviewing Schedule for Health Utilities Index (Mark III)

The following set of questions asks about each person's usual ability in certain areas, such as vision, hearing and speech. (Do not ask questions 2 to 32 for children less than 6 years old.)

1. INTERVIEWER CHECK ITEM:

Person 6 years or older →

Go to question 2

Person less than 6 years old →

Go to question 33

Vision

2. Are/Is _____ usually able to see well enough to read ordinary newsprint *without* glasses or contact lenses?

Yes → Go to 5

No

3. Are/Is _____ usually able to see well enough to read ordinary newsprint *with* glasses or contact lenses?

Yes → Go to 5

No

4. Are/Is _____ able to see at all?

Yes

No → Go to 7

5. Are/Is _____ able to see well enough to recognize a friend on the other side of the street *without* glasses or contact lenses?

Yes → Go to 7

No

6. Are/Is _____ usually able to see well enough to recognize a friend on the other side of the street *with* glasses or contact lenses?

Yes

No

Hearing

7. Are/Is _____ usually able to hear what is said in a group conversation with at least three other people *without* a hearing aid? Yes → Go to 12
 No
8. Are/Is _____ usually able to hear what is said in a group conversation with at least three other people *with* a hearing aid? Yes → Go to 10
 No
9. Are/Is _____ able to hear at all? Yes
 No → Go to 12
10. Are/Is _____ usually able to hear what is said in a conversation with one other person in a quiet room *without* a hearing aid? Yes → Go to 12
 No
11. Are/Is _____ usually able to hear what is said in a conversation with one other person in a quiet room *with* a hearing aid? Yes
 No

Speech

12. Are/Is _____ usually able to be understood completely when speaking with strangers? Yes → Go to 17
 No
13. Are/Is _____ able to be understood partially when speaking with strangers? Yes
 No
14. Are/Is _____ able to be understood completely when speaking with those who know _____ well? Yes → Go to 17
 No
15. Are/Is _____ able to be understood partially when speaking with those who know _____ well? Yes → Go to 17
 No
16. Are/Is _____ able to speak at all? Yes
 No

Getting Around

17. Are/Is _____ able to walk around the neighborhood without difficulty and without mechanical support such as braces, cane, or crutches? Yes → Go to 24
 No
18. Are/Is _____ able to walk at all? Yes
 No → Go to 21

19. Do/Does _____ require mechanical support such as braces, cane or crutches to be able to walk around the neighborhood?

- Yes
 No

20. Do/Does _____ require the help of another person to be able to walk?

- Yes
 No

21. Do/Does _____ require a wheelchair to get around?

- Yes
 No → Go to 24

22. How often do/does _____ use a wheelchair?

- Always
 Often
 Sometimes
 Never

23. Do/Does _____ need the help of another person to get around in the wheelchair?

- Yes
 No

Hands and Fingers

24. Do/Does _____ usually have the full use of two hands and ten fingers?

- Yes → Go to 28
 No

25. Do/Does _____ require the help of another person because of limitations in the use of hands or fingers?

- Yes
 No → Go to 27

26. Do/Does _____ require the help of another person with some tasks, most tasks, almost all tasks, or all tasks?

- Some tasks
 Most tasks
 Almost all tasks
 All tasks

27. Do/Does _____ require special equipment, for example, devices to assist in dressing because of limitation in the use of hands or fingers?

- Yes
 No

Feelings

28. Would you describe _____ as being usually:

(Mark one only)

- (a) happy and interested in life? a)
(b) somewhat happy? b)
(c) somewhat unhappy? c)
(d) unhappy with little interest in life? d)
(e) so unhappy that life is not worthwhile? e)

Memory

29. How would you describe _____ usual ability to remember things? (Mark one only)
Are/Is _____:
- (a) able to remember most things? a)
 - (b) somewhat forgetful? b)
 - (c) very forgetful? c)
 - (d) unable to remember anything at all? d)

Thinking

30. Would you describe _____ usual ability to think as: (Mark one only)
- (a) able to think clearly and solve problems? a)
 - (b) having a little difficulty when trying to think or solve problems? b)
 - (c) having some difficulty when trying to think or solve problems? c)
 - (d) having a great deal of difficulty when trying to think or solve problems? d)
 - (e) unable to think or solve any problems? e)

Pain and Discomfort

31. Are/Is _____ usually free of pain and discomfort? Yes → Finished
 No

32. Which one of the following sentences best describes the effect of the pain and discomfort _____ usually experiences? (Mark one only)
- a) pain and discomfort that does not prevent any activities? a)
 - b) pain and discomfort that prevents a few activities? b)
 - c) pain and discomfort that prevents some activities? c)
 - d) pain and discomfort that prevents most activities? d)

For information on the Ontario Health Survey (1990) contact:

Dave Bogart, Director
User Support Branch
Information & Systems Division
Ministry of Health Ontario
15 Overlea Boulevard
Toronto, Ontario M4H 1A9
Canada
Telephone: (416) 327-7610
FAX: (416) 327-7611

Larry Chambers, PhD
Department of Clinical Epidemiology
and Biostatistics
McMaster University
1200 Main Street, West
Hamilton, Ontario L8N 3Z5
Canada
Telephone: (416) 525-9140, Ext. 2136

C. QUALITY OF WELL-BEING SCALE AND GENERAL HEALTH POLICY MODEL

Contacts/Developers

Robert M. Kaplan, PhD, or John P. Anderson, PhD
Division of Health Care Sciences
Department of Community Medicine
School of Medicine, M-022
University of California, San Diego
La Jolla, California 92093
Telephone: (619) 534-6058
FAX: (619) 534-4642

Level	Definition of HRQOL concept	Preference weight
Mobility Scale (MOB)		
5	No limitations for health reasons	-0.000
4	Did not drive a car, health related; did not ride in a car as usual for age (younger than 15 years), health related	-0.062
3	Did not use public transportation, health related	-0.062
2	Had or would have used more help than usual for age to use public transportation, health related	-0.062
1	In hospital, health related	-0.090
Physical Activity Scale (PAC)		
4	No limitations for health reasons	-0.000
3	In wheelchair, moved or controlled movement of wheelchair without help from someone else	-0.060
2	Had trouble or did not try to lift, stoop, bend over, or use stairs or inclines, health related; limped, used a cane, crutches, or walker, health related; had any other physical limitation in walking, or did not try to talk as far or as fast as others the same age are able, health related	-0.060
1	In wheelchair, did not move or control the movement of wheelchair without help from someone else, or in bed, chair, or couch for most or all of the day, health related	-0.077

(continued)