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.

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.

*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 UM11C 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.

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 0.3–1.7.1 Research validating the power law has led to the conclusion that people can make remarkably consistent subjective
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].
DEBRA G. FROBERG and ROBERT L. KANE

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. 201.

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 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

\[ h_i = p + (1 - p)h_j \]

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
narrow 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 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

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

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

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
variable and determine what a particular scaling 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 utility 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>
<thead>
<tr>
<th>Study</th>
<th>SG</th>
<th>TTO</th>
<th>RS</th>
<th>ME</th>
<th>EQ</th>
<th>WTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patrick et al. [16]</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kaplan et al. [14]</td>
<td>No</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haig et al. [17]</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Torrance [38]</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wolfson et al. [39]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read et al. [34]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Llewellyn-Thomas et al [40]</td>
<td>No</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O'Connor et al. [36]</td>
<td>No</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detsky et al. [41]</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Churchill et al. [42]</td>
<td>No</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miles [43]</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

*A "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 Detinky, 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


