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.

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...
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, 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 nonresidents; 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.

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 intra-rater 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 expected utility model does not adequately describe decision-making in situations of social isolation, as in research conditions. This makes it difficult to predict new context effects. Further, the problem representation and will therefore not easily predict new context effects. 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 Detaksy, Walter Spitser, 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