Validity and Interpretation of Preference-Based Measures of Health-Related Quality of Life

Leslie Lenert, MD, MS,*† and Robert M. Kaplan, PhD‡

Utilities are numeric measurements that reflect an individual’s beliefs about the desirability of a health condition, willingness to take risks to gain health benefits, and preferences for time. This report discusses the approaches to assess and compare the validity of methods used to assign utilities for cost-utility analysis. Threats to validity include construct underrepresentation and construct-irrelevant variance. Construct underrepresentation occurs when a stimulus presented to a judge fails to fully represent the depth and complexity of information required in actual judgments. Construct-irrelevant variation occurs when factors irrelevant to preferences influence measurements of utilities. Among several factors that cause construct-irrelevant variation are cognitive abilities, numeracy skills, emotions and prejudices, and the elicitation procedure. Commonly used elicitation methods (visual-analog scales, time tradeoff, and standard gamble) capture different facets of utilities (desirability of states, time preferences, and risk attitude) to different degrees. The validity of an elicitation protocol depends (1) on the degree to which its scaling method captures the relevant facets of utility and (2) on the degree to which measurements are influenced by construct-irrelevant variation. Discrete-state health index models provide an alternative to direct elicitation of utilities and work by attaching fixed preference weights to observable health states. The creation of discrete-state models with current technologies requires the adoption of strong assumptions about the scaling properties of utilities. Future research must refine methods of eliciting utilities and identify sources of construct-irrelevant variability that reduce the validity of utility assessments. Because of the impact of variation in techniques on measurements, we do not recommend the combination of utilities elicited with different protocols in cost-utility analysis and do not recommend the display of cost-utility ratios from different studies in comparison or “league” tables.

Key words: Quality-adjusted life-years; health surveys; psychometrics; questionnaires; utility theory; standard gamble; time tradeoff. (Med Care 2000;38[suppl II]:II-138–II-150)

The use of health-related quality of life (HRQOL) measures in the evaluation of health care has increased dramatically during the past decade.¹ Generic measures of HRQOL can be divided into 2 broad categories: psychometric profiles and decision-based utility measures. The psychometric measures, best exemplified by the Medical Outcomes Study 36-item Short Form (SF-36), typically offer a profile of health outcomes.² For example, the SF-36 includes 8 health concepts: physical activity, role—physical, bodily pain, general health perceptions, vitality, social functioning, role—emotional, and mental health. Psychometric profiles provide data on functioning relative to both a minimal level of performance for each health concept (eg, a floor) and a maximal level (eg, a ceiling). If calibrated to a population standard, scale scores also provide information on the
degree of abnormality of an impairment of health for a group or an individual, but they offer little information on the significance of the impairment.

Utility-based measures were developed to address the issue of the significance of health impairments in a systematic way. These measures share some similarities with psychometric measures of health. For example, both measures often examine the same dimensions of health. However, although psychometric measures report a series of scores and characterize respondents on each independent dimension of the profile, utility measures use human judgment to combine and scale health effects over several different dimensions. The scaling of utility measures is always made in terms of some absolute reference point (often, “perfect health” and death), as opposed to the population reference point used in many health-status measures.

Validity of Preference-Based Measures

This report explores the issue of the validity of preference measures and considers the problem of the comparison of results across studies that use different protocols. Definitions of validity and reliability are taken from the 1999 edition of Standards for Educational and Psychological Tests. Recommendations in this book were developed through a consensus process with participants from the American Educational Research Association, the American Psychological Association, and the National Council in Education in Measurement. The recommendations were designed to promote the sound and ethical use of psychological as well as other types of tests. In this book, the validity of a test or measure is defined as the degree to which evidence and theory support the intended interpretations of a test or a measure. The authors of the consensus statement argue that validity should be considered a unitary concept that cannot be separated into components or discussed in isolation from the intended interpretation of a measure. They caution against the separation of validity into subcategories such as content validity, predictive validity, and criterion validity. Content, criterion, and predictive evidence might support particular interpretations of test scores but are not conceptually independent from the intended interpretation of the test. The latest version of the standards emphasizes 2 new concepts: construct-irrelevant variance and construct underrepresentation. Construct-irrelevant variation occurs when scores are influenced by factors irrelevant to the construct. Construct underrepresentation describes the failure to capture important components of the construct. This report applies this new conceptual framework to characterize the validity of measurements of utilities.

Our discussion focuses on the validity of utility and preference measures for use in cost-effectiveness analysis as weights for life-years (eg, cost-utility analysis). In this context, the characterization of the validity of a preference assessment method might be thought of as an assessment of the plausibility that a chain of evidence that links a population preference or utility weight to preferences of a relevant group or society as a whole. There is some debate about whether utilities exist or are created in the process of measurement. For the present purposes, we assume that there is a number, which we will call a “utility,” that if measured successfully, would predict an individual’s real-world choices. The construct of utility has 2 elements: the health condition or state and the desirableness of that condition or state to the individual. Assessments of construct-irrelevant variation and construct underrepresentation and of overall validity for use should address both of these elements.

In the measurement of preferences for cost-effectiveness analysis or guideline development, the goal is to measure preferences for a health condition in a relevant group and then to aggregate individual measurements into a single numeric weight that reflects the overall desirableness of the health condition. In practice, utility assessment is performed with varying degrees of success. For some individuals, measurements mirror the values that underlie real-world decision processes. For others, measurements are largely artifacts of the assessment process that are shaped almost entirely by causes of construct-irrelevant variation. Because it is difficult to use future behavior as a standard by which to judge the validity of utility measurements methods, one of the critical issues that we discuss here is how to assess the likely degree of correspondence between a measurement result and an individual’s “true” preferences.

We propose that researchers and policy makers take a bayesian approach to interpretation of utility data, as illustrated in Fig. 1. Before measurement, the best guess at an individual’s utility (eg,
Utilities, Construct-Irrelevant Variation, and Construct Underrepresentation

The success or failure of a preference measurement is the result of a number of factors, as illustrated in Fig. 2. Preference ratings, even for current health, are formed in response to a stimulus that prompts the subject to think about the state. As shown in Fig. 2, subjects must process information from health state descriptions to provide judgments. A critical factor here is the adequacy of descriptions of the state to be rated. This stimulus is typically text based but sometimes has included voice recordings, pictures, or brief video clips. The use of multimedia methods may facilitate the processing of this information. Subjects who view multimedia presentations have better recall recognition and better integration of materials into their preferences. In addition, cognitive skills are required to translate descriptions into assessments of the implications of health impairments. Health impairments further often impair this ability and thus are a frequent cause of construct-irrelevant variation. The rating task is also dependent on a subject’s numeracy or quantitative reasoning skills. If subjects have little experience with quantitative reasoning, they may perform quantitative rating tasks poorly. Consequently, the specific numeric values elicited in an experiment may not be closely connected with true preferences.

Another way in which utility elicitations can fail is if the material in a health state description triggers an emotional response that impairs judgment. Pictures or text descriptions of states may trigger hidden prejudices or fears that are irrelevant to values for the state, resulting in construct-irrelevant variation. Some content may produce emotional arousal that could degrade numeracy skills, leading to random errors. Other content may spuriously influence preferences for the state described. For example, Lenert et al found that subjects who rated states in which health conditions were depicted by an actor of a different race assigned lower utilities to states than did subjects when the actor was the same race. Patrick et al anticipated such problems and designed generic stimuli that explicitly excluded information about race, gender, and ethnicity. However, a retreat to entirely bland stimuli may not be desirable in that this may lead to a loss of validity due to construct underrepresentation.

Several facets of utility are necessary in measurements of preferences to avoid the loss of validity due to construct underrepresentation. As shown in the figure, these include the desirability of the depicted state, risk attitude (or willingness to gamble to gain health benefits), and preferences for sequences of events or time. Different scaling methods capture different aspects of preferences, as discussed later. Each scaling method produces measurements that omit certain facets of utility. Furthermore, construct-irrelevant variation may lead to failure of the utility assessment process (eg, a measurement not predictive of an individual’s behavior). The assessment process in turn may fail
for several different reasons. Each mode may have a different type of effect on the end value observed. Effects of variation due to scaling methods, procedures for assessment, and failure of the elicitation process are discussed as follows.

**Construct-Relevant Variation in Preferences**

Not all variability in preferences will be the result of spurious influences on measurements.

---

**Fig. 1.** Model illustrating factors underlying a population preference weight. The population weight is a summary of the population distribution of preferences. This in turn is a summary of the posterior (e.g., postassessment) distributions of individuals’ true preferences.

---

**Fig. 2.** Model of the process of obtaining a preference elicitation.
Valid measures of preferences must capture the genuine variability in preferences among individuals and groups in society. The identification of such differences in preferences between groups in society is of particular relevance in cost-utility analysis. A particular group might have health preferences that are systematically different from those of society as a whole because their needs and desires are different. Consequently, the allocation of health resources with societal preferences might lead to discrimination against this particular group.¹² Not all differences in values would result in discrimination; some might result in a preferential bias or have no effect.¹²,¹³

One group of individuals who may have systematically different preferences from society as a whole are patients. Several studies have found evidence of such differences.¹⁴–¹⁶ Recent work has also documented differences in preferences between family members of patients and the general public.¹⁷ Findings of differences in preferences between patients and other members of society are by no means universal in studies, however.¹⁸ Recent work has focused on the issues that may underlie the variance in results of studies, such as whether present health status or past experience with disease drives differences in preferences.¹³,¹⁵,¹⁶ There also is some evidence that ethnicity of patients is associated with systematic differences in preferences.¹⁹,²⁰ Although controversy remains about whether there are differences in preferences between patients and the public and among ethnic groups, if these differences do exist, valid measures of preferences should reveal, not obscure, these differences.

Scaling Methods and Construct Underrepresentation and Construct-Irrelevant Variation

At the end of the process elicitation depicted in Fig. 2, scaling methods are among the most important aspects of measurement of utilities. Methods for assessment of utilities have been discussed extensively in other publications.¹² Von Neumann and Morgenstern²¹ outlined axioms of choice that have become basic foundations of measurement of utilities. This work was extended by Raiffa²² and several others (see reviews by Bell and Farquhar,²³ Howard,²⁴ and Torrance and Feeney²⁵).

The best known method for the assessment of utilities is the standard gamble (SG). With this technique, a respondent is given a hypothetical choice between continued life in a current state of health or a gamble that would result in perfect health (with a probability of \( P \)) or death (with a probability of \( 1 - P \)). The SG is appealing because it allows an assignment of utility and directly incorporates risk aversion into the decision process.

An alternative method is the time tradeoff (TTO), in which respondents are asked the amount of time they would be willing to give up to be in a better health state.²⁶ The major advantage of the TTO is that responses are conceptually linked to the quality-adjusted life-year model; therefore, the raw data require little manipulation for incorporation into cost-effectiveness analyses. Many researchers consider the TTO easier to implement in clinical studies than the SG, although there is no published evidence to support this claim.

A third approach involves the use of simple rating scales (RSs). A common variant of this approach is a visual-analog scale. With this method, subjects are required to rate health conditions on a scale that ranges from 0 to 10 or from 0 to 100. Ideally, the end anchors are clearly defined, with 0 equal to death and 100 (or 10) equal to perfect health. Scales are typically oriented vertically rather than horizontally. Unlike SG and TTO, subjects are not required to make a choice between alternatives. In addition, RSs do not consider attitude toward risk or incorporate time horizons.

Comparisons among SG, TTO, and RSs typically show that the methods yield different preference weights.²⁷–³¹ Usually, preferences obtained with the SG are higher than those obtained with the TTO. In turn, TTO preferences are usually higher than those measured with an RS.³² However, these findings are not always consistent because the effects of changes in procedures for measurement (search method for indifference points, graphical feedback on tradeoffs, and so on, as discussed later) can be as large as the effects of assessment methods.³³

The “correct” assessment method remains a matter of debate. SG is often advocated because of its close ties to the theoretical foundations of concepts of utility.²⁵,³⁴ Richardson³⁵ has argued that TTO should be used as a validity criterion. Broome³⁶ has argued that SG and TTO are too restrictive in terms of individual preferences, so RS should be the preferred method. Rarely is it rec-
Recognized that the 3 methods ask different questions, so each method has its own focus. Visual-analog scales primarily measure the severity of a health condition relative to anchoring stimuli (typically, perfect health and death). This procedure also measures a subject’s risk attitude (willingness to gamble to achieve health benefits). Given the differences in what is being measured, it is not surprising that most investigators have found large differences in the specific numeric values obtained. Because each of these scaling methods assesses different aspects of preferences, the degree to which construct underrepresentation is a problem depends on the aspects of preferences that a researcher considers to be relevant for inclusion in a cost-effectiveness analysis. Thus, if a researcher or policy-maker deems that risk attitude is a vital aspect of preferences for consideration in a model (and the results of decision analyses and cost-effectiveness analyses can be sensitive to risk attitude), measurement with the SG is the only appropriate alternative. If time preference is deemed to be of substantial importance, then TTO (or a gamble with a specific duration of life) may be a more appropriate tool than a visual-analog scale. If the objective is comparison of the perceived overall severity, then analog scale ratings may be the preferred metric (because other aspects of preferences may least influence results obtained with this method).

Protocol Implementation and Construct-Irrelevant Variation

The completion of a preference-elicitation protocol does not mean that the measurement value obtained is an unbiased and accurate representation of an individual’s utilities. As shown in Fig. 2, the degree of correspondence between measurements of values and utilities may be influenced by health state description methods, cognition, numeracy, assessment procedures, and interactions among these factors.

Construct-irrelevant variation due to procedures for elicitation is one of the important limitations of preference measurement. A change in the procedures for measurement can alter the results obtained with all 3 scaling methods. Analog scaling methods are prone to sequencing effects. The specific values obtained with RSs can influence the appearance of the scale (eg, an open-ended question versus a horizontal scale rating versus a vertical scale), the range of health states considered during the performance of preference ratings, and the prior assessments performed with other methods. Utilities for the SG and TTO can be influenced by a number of procedural factors, such as the frame of the gamble (with a focus on positive or negative outcomes) and the bottom anchor in the gamble. Both the SG and the TTO may be influenced by the specific procedure used to find a subject’s indifference point. In the SG, the specific probabilities that are used may influence results due to the effects of the prospect theory. In addition, the TTO requires subjects to contemplate a specific duration of length of life for the base case. Values obtained with the TTO are influenced by the duration of survival in the base case. Under severe conditions, patients often refuse to tradeoff length of life because they place a greater value on survival than on the quality of life. If the condition is particularly painful or unpleasant, patients may express nonmonotonic preferences for survival duration.

Because of construct-irrelevant variation, the results of assessments performed with the same scaling methods but with different protocols for assessment are not directly comparable without an accounting for differences in procedures. A detailed examination of assessment protocols is needed to be able to determine the comparability. Documentation in the published literature of the procedures that use utility assessment protocols is often inadequate to assess the comparability of different protocols for these reasons.

Measurement of the Effects of Construct-Irrelevant Variation and Construct Underrepresentation

In this section, we discuss how researchers can assess the validity of a preference measurement protocol by examining group- and individual-level data for evidence of failure of the elicitation process. One important mode of failure is low reliability due to random error in assessments. In psychometric research, random error is typically measured in studies of the test-retest reliability of an instrument. However, characterization of the reliability of preference elicitations may not be as simple as it is for psychometric measures. Psychometric theory presumes that behaviors or preferences are stable over time. A critical issue is whether alterations in responses at any 2 points in...
time are the result of random error. An additional factor in many protocols is the use of a human interviewer who presents material to subjects and actively participates in the assessment process. The reliability of responses can be both positively and negatively influenced by nonverbal cues that interviewers provide to subjects during the assessment process. Furthermore, not all changes in responses from test to retest are the result of random noise in measurement. Some changes may be due to an improvement in a subject’s ability to express his or her preferences within an elicitation framework with repetition.33

Current approaches to the assessment of random error in utility measurements are based on calculation of the intraclass correlation coefficient. Conceptually, this model excludes fixed time-to-time variation and focuses on random components. However, this excludes the possibility of changes in preferences or failure by nonrandom error. If utilities take only a few specific values because of risk aversion, unwillingness to trade off time, or underrepresentation of the health state construct, this can lead to an underestimation of random error. For example, published data from the Beaver Dam Health Outcomes Study,47 a large cross-sectional survey of utilities for current health and common diseases, show how utility assessment procedures can change a continuous measure into a categorical measure. In this study, >80% of preference ratings took 1 of 3 values: 1.0, 0.9, or 0.5. When the preponderance of utility values in a protocol take 1 of a few values, Kiebert et al48 suggested that reliability be measured with Cohen’s κ. However, excessive imprecision in measurement suggests that a protocol is not useful as a tool for the measurement of the preferences of individuals. Furthermore, it is likely that population mean values derived from such a protocol would be biased. The tendency for utility assessments to take a few discrete values in some protocols suggests that this is an important mode of failure of the utility elicitation process. Several authors have proposed the use of an individual’s ability to resolve his or her preferences for states as a benchmark of the validity of an elicitation protocol.49,50 Researchers may be able to foster “resolution” in utility measurements by describing health states in a manner that promotes strongly held preferences and by a careful search for a subject’s indifference point in elicitation protocols.

Perhaps the most important evidence that utility elicitations are excessively affected by construct-independent variation or construct underrepresentation is a lack of internal consistency in preference measurements. The presence of occasional inconsistencies in health utilities does not indicate that a procedure for preference assessment is invalid. Rather, it indicates that in some proportion of subjects, the procedure was unsuccessful in eliciting their true preferences. Two types of errors in internal consistency have been described. One type of error is an illogical ordering of health states within any one given assessment method.51 The other type of error is an inconsistency in the rank ordering of preferences for states across different methods of assessment52 (also called failure to satisfy procedural invariance).9

The identification of a logical type of error requires a set of health states that have inherent canonical relationships with each other. On the basis of these relationships, one should be able to infer that states with lesser levels of impairment should be preferable to states with greater levels of impairment or with additional impairments in other aspects of health. Dolan and Kind53 compared the validity of different protocols based on the rate of such errors. Rutten-van Molken et al51 attributed logical misorderings to random error in the preference-elicitation procedure. If this is true, most subjects should choose to correct the misordering if they are informed of it. Lee et al17 confirmed that subjects correct misorderings of states when presented with computer reminders of errors within assessments.

The identification of logical errors in preference elicitations requires a set of states with a defined rank order of desirability. Although this is not an uncommon structure in utility elicitation experiments, it is by no means a universal one. To assess the internal consistency of utility elicitations when there is no logical order among states, Lenert et al52 proposed a comparison of the rank order preference assessments across different scaling method. Conceptually, any 2 states (a and b) can have 1 of 3 cardinal relationships: a > b, a = b, or a < b. Rational persons should satisfy procedural invariance—that is, they should have the same relationship among 2 states among measurements performed with any given scaling method. An observation that 2 states are judged equally desirable with 1 method but not the other (eg, a1 > b1 and a2 = b2) suggests a possible failure to discriminate among states but not necessarily a violation of procedural invariance.9
Comparisons between the preferences of violators and satisfiers of procedural invariance suggest that these groups are 2 distinct populations. Violators often have preference measurements that take different mean values than do satisfiers. Violators also appear to have less ability to discriminate between states and less organization among preference ratings for a series of states. At best, measurements of preferences are less successful in violators; at worst, measurements may be mere artifacts created by the procedures used for preference assessment.

Giesler et al. suggested an alternative approach for the assessment of internal consistency that uses a subject's rank ordering of preferences for a series of states as a pseudo gold standard for the evaluation of internal consistency. However, there are no data to suggest that rank ordering of states is performed with greater or lesser accuracy than are more quantitative procedures for assessment. If the rank ordering procedure is not assumed to have special properties with regard to ascertainment of the true order of preferences for states, the approach contributes both conceptually and methodologically to the assessment of procedural invariance.

Subjective Measures of Effects of Construct-Irrelevant Variation and Construct Underrepresentation

Several investigators have used subjective approaches to measure the overall degree of correspondence between measurements and an individual’s utilities. The approach applied in Lenert and Soetikno, Bergen et al., and Lenert et al. characterizes validity as a preference assessment protocol based on the judge’s beliefs about the properties of his or her own measurements. This framework breaks down self-assessments of validity into 3 constructs: beliefs about the ease of use, self-reported confusion about the meaning of the reported value, and comfort in use of the procedure in future decision making. Each of these attributes is scored on a 4-point Likert scale. Other investigators have proposed having assessors quantify the validity of the elicitations they obtain from subjects through the use of a single global Likert scale. Such subjective measures of validity of elicitations appear to assess different aspects of validity of preferences than do objective measures. Lenert et al. have found that judges’ ratings of their own understanding of preference assessment procedures were neither sensitive nor specific predictors of inconsistencies in preference ratings.

Validity, Construct Underrepresentation, and Intended Use

Construct-irrelevant variation and construct underrepresentation are powerful tools that explain why widely used procedures for assessment can produce invalid results in specific contexts. However, there are other factors that need to be incorporated into assessments of the validity of use of utilities as societal preference weights. Societal decision values often include other factors than preference values for states, including concerns for the fairness (equity) of distribution of health benefits, for providing help to the most needy, for maintaining hope, and other factors. One unanswered question is whether the current paradigm of aggregation of individual utilities can yield valid societal preference weights. To be valid representations of societal preferences, preferences may need to be elicited in the context of ex ante social decision making, with approaches such as the person tradeoff or willingness to pay for insurance premiums. The absence of such elements from the frame of utility elicitation might result in construct underrepresentation. Alternatively, researchers could adjust aggregated individual preference weights to incorporate such factors. Nord et al. proposed such adjustment methods. In addition, 1 potential measure of the validity of a utility elicitation protocol is the degree to which measurements capture these other aspects of societal values without the need for ad hoc factors to adjust preference ratings. For example, the use of scaling methods that take risk attitude into account may result in recommendations that are more consistent with concerns about fairness and help for the most severely ill and about the necessity for hope.

Construct-Irrelevant Variance and Health Index Models

An alternative approach for the estimation of utilities in clinical trials and cross-sectional studies...
is the health index model. One important advantage of health index models such as the EQ-5D, Health Utilities Index (HUI), and the Quality of Well-Being (QWB) scale is that these models use a single consistent approach to measure preferences for states in the model. Therefore, the results from analyses are easier to interpret. However, in health index models, the chain between population numeric value and individual measurements of preferences has an additional link that may affect the validity of results through the introduction of construct-irrelevant variance. Rather than attempt to find utilities for specific conditions, the developers of health index models first define a general discrete-state health status model. They then measure utilities for some of the states in the model and estimate utilities for the remainder with statistical procedures. To use a health index model in a clinical study, a researcher determines which of the many discrete states in the model correspond to a patient’s current state of health. The researcher then looks up the utility of the state and assigns that value to the patient.

Health index models examine multiple dimensions of health or attributes of health such as pain and lack of mobility. The exact dimensions vary from measure to measure. There is considerable debate about what dimensions should be included in health index models. Most measures include physical functioning, role functioning, and mental health. Attributes are divided into multiple levels of functioning. The resulting models are complex, with thousands of possible combinations of levels of attributes. The HUI Mark III is an extreme example, with 972,000 discrete states. The number of states in index models precludes the direct measurement of values for all combinations of different levels of attributes. Recent work has shown that k-means clustering to identify the states that exist in nature can reduce the number of states required to represent variability in functioning.

Although the assignment of a utility score in a patient does not require judgment from the patient in a clinical study, the creation of these models requires the same sort of global judgments of the overall quality of life for a case as previously discussed. Typically, to populate health index models with preference data, researchers conduct large-scale population surveys. By systematically varying the attributions in cases, researchers can estimate population weights for the effects on each dimension of health on preferences. Although the studies conducted to populate health index models with preference data are large (typically thousands of subjects), in most models, the number of states far exceeds the number of subjects in the study. As a result, relatively strong assumptions about how dimensions represented in the model interact have been required to complete the model. This degrades the validity of utility estimates by adding additional construct-irrelevant variance.

Assumptions about interactions are a problem primarily if they are inconsistent with the process that humans use to form integrated judgments of multiattribute problems. Substantial evidence suggests that human judges most often integrate multiattribute information with an averaging rule. The averaging rule yields an additive model of human judgment. This averaging process has been validated in specific experimental tests (see Anderson [Volume III] for a review of evidence). This finding implies an additive utility function and is consistent with preference models with mutual additive utility independence. Specifically, mutual additive utility independence assumes that there is no preference interaction among the attributes being rated. In other words, overall preference scores depend on the individual level attributes and not on the unique combinations between attributes.

Under assumptions of mutual additive utility independence, the task of estimating preferences for health states is greatly simplified, as is the task of applying the model to other populations. Tests of mutual additive utility independence among preference measurements take 2 forms. As illustrated in Fig. 3, additive utility independence is present if the decrement in utility for a given disability is constant across different levels of a second attribute. The second test involves an examination of the “corner states” of utilities (states with full functioning in all except 1 level of health and minimal functioning in the other level). Under conditions of mutual utility independence, utilities for corner states should sum to 1.0. Measurement of the conformance of model predictions with additive utility independence assumes assessments arise from a scale with interval properties. Nonconformance could be due to either the measurement scale not having interval properties or preferences having a more complex structure (or both).
There are conflicting data on the conformance of preferences to the assumption of mutual additive utility independence. The San Diego group and the Oregon Medicaid Experiment have found evidence for mutual additive utility independence in measurements for the QWB. However, the use of tests of utility independence performed with other health index models (eg, those performed by Torrance et al, Dolan, Krabbe et al, and, most recently, Hakim and Pathek) has been unable to confirm this finding. This result is not surprising, because the dimensions represented in the models and the processes used to elicit preferences differ.

Differences Between Health Index Models and Other Models of Utility

A critical issue is the degree to which utilities from general models capture the full burden of the effects of a specific disease on quality of life. To the extent that there is variation in the degree to which a model captures the burden of disease, there is the potential for bias. One approach to assessment of the potential bias from the use of health index model is to compare changes in utilities estimated through disease-specific methods or longitudinal intrapersonal measurements with changes estimated from the application of health index models.

Conclusions and Recommendations

Both “validity” and “utility” are complex constructs. “Validity” defines the range of inferences that can be made on the basis of a score or measure and limits what can be said or should not be said. Our review suggests that there has been significant progress in theory and methodology for utility assessment. On the other hand, the inferences derived from the application of utility measures should not exceed what the validity of these measures can support. Our recommendations for the use of preference weights in cost-utility analyses follow:

![Graph showing QWB Rating vs. Physical Activity]

**FIG. 3.** Under conditions of mutual additive utility independence and measurement of utilities with a metric with interval scaling properties, differences between effects of a change in health in 1 dimension or attribute are independent of each other.
1. A major source of variability in utility ratings results from the use of different scaling methods in measurement. RSs, TTO methods, and the SG arise from different theoretical constructs and ask different questions. Each method can potentially fail due to construct underrepresentation when applied to specific clinical conditions, and no one method is theoretically “correct” in all circumstances. Researchers should choose an assessment method that adequately represents facets of utility relevant to the problem area.

2. The utility elicitation is a process that is not always successful for every individual. Therefore, it is important that researchers take a bayesian interpretation of preference measurements and weight preference ratings in proportion to the evidence for the validity of the measurement or lack thereof.

3. In addition to sporadic failures of the elicitation process, construct-irrelevant variance and construct underrepresentation cause preference measurements to be sensitive to both the choice of scaling method and the procedures for implementation. The combined effects of several alterations of procedures on measurements are difficult to predict, and there are relatively few studies that explore this issue.

4. The procedural sensitivity of elicitations has important implications for the use of preference measurements in cost-utility analysis. The level of documentation in most published articles is not sufficient to appraise how alterations in elicitation procedures may bias results, even if the effects of specific combinations of changes in procedures are known. For this reason, we recommend against the combination of utilities elicited through the use of different protocols in cost-utility analyses and against the display of cost-utility ratios in league tables that summarize analyses from multiple studies. When researchers must combine utility measurements taken from studies that used different protocols, we recommend that they subject the results to extensive 2-, 3-, or even n-way sensitivity analyses to utilities. All league tables should include cautionary advice to the reader.

The procedural sensitivity of preference measurements demands that greater efforts be directed toward the development of standardized utility elicitation protocols. Standards should be based on empirical studies that compare the reliability of measurements, subjects’ ability to discriminate among states, and the internal consistency of different protocols. One approach to facilitate standardization is the greater adoption of computerized methods to conduct preference-elicitation interviews. Software to construct computerized utility-elicitation protocols is available to researchers from several sites on the Internet for free or at a low cost, including the Web site of one of the present authors (http://preferences.ucsd.edu).

In addition, we urge that more research be conducted on the validity of utilities. The best currently available strategy to assess or compare validity of a protocol or protocols is to examine the results of elicitation for evidence of effects of concept-irrelevant variation and concept underrepresentation. However, future work should also examine linkages between preference measurements and actions of judges and should compare the validity of protocols based on their ability to predict the judges’ real-world behaviors with measurements of preferences.

Acknowledgments

This work was supported in part by grant R01-HS-09170 from the AHCPR, grant P60-AR-40770 from the NIH, grant TPRH-98-119-01 from the American Cancer Society (Dr Kaplan), and grants R01-AR-46125R and R29-LM-05626 from the NIH (Dr Lenert).

References


VALIDITY AND INTERPRETATION OF UTILITIES


