Invariance and Inconsistency in Utility Ratings

*Dena M. Bravata, MD, MS, Lorene M. Nelson, PhD,
Alan M. Garber, MD, PhD, Mary K. Goldstein, MD, MS*

Purpose**. To assess utilities of composite health states for dependence in activities of daily living (ADLs) for* invariance *(i.e., when subjects provide a utility of 1 for all health states) and* order inconsistency *(i.e., when subjects order their utilities such that their utility for a combination of ADL dependencies is greater than their utility for any subset of the combination).* ***Methods**. Each of the 400 subjects, age 65 y and older, enrolled in one of several regional medical centers of the Kaiser Permanente Medical Care Program of Northern California and provided standard-gamble utilities for single ADL dependencies (e.g., bathing, dressing, continence) and for dependence in 8 other combinations of ADL dependencies. For order-inconsistent responses, the authors calculated the maximum magnitude of inconsistency as the maximum difference between the utility for the combined ADL dependence health state and that of its inconsistent subset.* ***Results**. A total of 76 subjects (19%) gave a utility of 1.0 for all health states presented to them; 19 (5%) gave the same utility other than 1.0 for all health states; 130 (33%) gave at least 1 utility < 1.0 and had no order inconsistencies; and 175 (44%) had at least 1 order inconsistency. Invariance was associated with a Mini-Mental Status Examination score < 28.6 (*P *= 0.01), with education < 12 y (*P *= 0.004), with race/ethnicity other than non-Hispanic White/Caucasian (*P *= 0.001), and with shorter time spent on the utility elicitation task (*P *< 0.0001). Among the inconsistent subjects, 69% had a maximal magnitude of inconsistency that was within 1 standard deviation of the mean utilities. The maximal magnitude of inconsistency was associated with longer time spent on the elicitation task (*P *< 0.0001) and race/ethnicity other than non-Hispanic White/Caucasian (*P *= 0.005). The mean (s) utility for dependence in continence among consistent subjects who were not invariant (0.88 [0.24]) was higher than among inconsistent subjects (0.80 [0.27];* P *= 0.01).* ***Conclusions**. Invariance and order inconsistencies in utility ratings for complex health states occur frequently. Utilities of consistent subjects may differ from those of inconsistent subjects. Utility assessments should attempt to measure and report these patterns.* ***Key words: *activities of daily living; quality of life; utility theory.* ***(Med Decis Making 2005;25:158–167)***

There has been considerable effort to develop methods to minimize inconsistencies in patient preferences that result from interviewer bias, respondent fatigue, framing effects, order effects, and search procedures.[1–5] However, there is no clear consensus as to the best method for comprehensively evaluating inconsistencies in patient preferences, for assessing the populations most at risk for inconsistencies, or for managing inconsistent responses when they are found.

The literature describes 3 methods for evaluating inconsistencies in assessed utilities. First, inconsistency has been measured as the degree to which a respondent's reported utilities deviate from expected ordering (e.g., a subject may be expected to prefer the state of monocular blindness to binocular blindness).[6–8] Rates of these order inconsistencies range widely from 17% to 49% of subjects asked to rank multiple health states, depending on the clinical condition and subjects surveyed.[9–11] Second, some researchers have evaluated consistency by asking subjects to rank the outcomes of interest prior to assessing preferences and then compared reported utilities with this ranking (e.g., if outcome A is ranked above outcome B, then the utility for A should be greater than the utility for B).[1,10,11] Finally,

Received 28 January 2003 from the Center for Primary Care and Outcomes Research, Department of Medicine, Stanford University School of Medicine, Stanford, California (DMB, AMG, MKG); Division of Epidemiology, Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California (LMN); VA Palo Alto Healthcare System, Palo Alto, California (AMG, MKG); and Division of Health Services Research, Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California (MKG). Portions of this work have been presented at the Annual Meeting of the American Geriatrics Society, Nashville, TN, May 2000, and at the 22nd Annual Meeting of the Society for Medical Decision Making, Cincinnati, OH, September 2000. The abstracts for these presentations can be found in *Journal of the American Geriatrics Society* 2000;48(8):S117 and *Medical Decision Making* 2000;20(4):489. Revision accepted for publication 23 January 2005.

Address correspondence to Dena M. Bravata, Center for Primary Care and Outcomes Research, Stanford University School of Medicine, 117 Encina Commons, Stanford, CA 94305-6019; fax: 650-723-1919; e-mail: dbravata@stanford.edu.

DOI: 10.1177/0272989X05275399

others have attempted to verify that subjects consistently order utilities across methods of preference assessment.[1,3,12–14]

Another form of potentially invalid utility responses is invariance, where a subject assigns the same utility value to all states assessed. Among preference assessments that include more than 1 health state, researchers have noted that some subjects assign the same utility rating to all health states. Although there is little literature on invariance, Rutten-van Molken and colleagues found that 8 of 85 (9%) patients with fibromyalgia and 11 of 144 (8%) patients with ankylosing spondylitis gave the same rating (0.95) to all 3 health states evaluated.[8,15]

Order and other logical inconsistencies have been associated with respondent characteristics and factors related to the preference elicitation. Specifically, inconsistencies tend to be more common in respondents who are older and those who have poor numeracy, cognitive defects, worse health status, or less education.[9,16–18] Inconsistencies are also more common when interviewers have little experience in preference assessment and when the standard-gamble and time-tradeoff methods are used as compared with the rating scale.[1,9–11,19]

Three approaches have been reported to "repair" or manage inconsistent responses. First, some interviewers (and computer-based elicitation programs) allow subjects to change their stated preferences as their understanding of the issues deepens.[9,20] Second, some computer-based elicitation programs will not allow subjects to give a rating for a severe health state that is better than a previously reported rating for a milder health state and vice versa.[9] Finally, some researchers eliminate the utilities of respondents whom the interviewer viewed as having poor comprehension of the utility task and of respondents whose utilities differed markedly from the mean or median of the population studied.[21,22] For example, 74 of 293 respondents in the Health Utilities Index Mark 2 study were excluded for illogical rankings.[23,24] Excluding inconsistent responses may alter the mean utility for a health state if the preferences of those excluded differ from other population subgroups or from the general mean. A consensus report on what should be included in the methods section of a paper reporting utilities has been published[25]; however, there is no consensus statement on how to handle utilities from respondents who were inconsistent or invariant. Available reports on inconsistency and invariance are scant; additional data are needed that explore these response types and can be used to inform a plan for managing them.

We sought to evaluate the rates of both invariant responses and inconsistent ordering among elderly subjects providing standard-gamble utilities for health states of single and combined dependencies in the personal self-care capacities known as activities of daily living (ADLs; eating, bathing, dressing, continence, using a toilet, and transferring in and out of a bed or chair).[26,27] We hypothesized that the frequency of both invariant and inconsistent responses would be more common among respondents who were older, had more medical problems, and had lower educational and socioeconomic status. Finally, we attempted to evaluate differences in utilities between consistent and inconsistent respondents and to explore how including and excluding invariant and inconsistent responses affect mean utilities.

MATERIALS AND METHODS

Subject Recruitment

Eligible subjects were patients age 65 y or older enrolled in one of several regional medical centers of the Kaiser Permanente Medical Care Program of Northern California. We wrote and then called 1382 randomly selected subjects to verify eligibility and to recruit suitable patients for the study. A total of 201 subjects were excluded because they were deceased or had left Kaiser, could not speak English, had a terminal illness, or had significantly impaired cognition, sight, or hearing. An additional 75 subjects were excluded because their contact information was not up to date or because they could not be reached by telephone after many attempts. Of those who were eligible and able to be contacted, 661 did not elect to participate. For those eligible and willing to participate, we offered to conduct the interview at a time and place convenient to subjects, whether at their homes, in a nearby Kaiser facility, or at our offices. Subjects were paid $20, and each provided informed consent. Our protocol received Human Subjects' Committee approval from Stanford University.

Data Elicitation

A research assistant interviewed each subject to obtain demographic information and scores on the Mini-Mental State Examination (MMSE).[27] Subjects with scores of 23 or lower on the MMSE, indicative of cognitive impairment, were excluded. The remaining subjects were interviewed with the computer utility elicitation program Functional Limitations and Independency Ratings, version 1 (FLAIR1).[28] FLAIR1 begins with a general introduction followed by training and practice in use of the pointer, trackball, and the buttons on the screen.[28–30] Subjects are then introduced to the general concept of dependency in ADLs, provided with a description of each ADL, and asked to classify themselves as either needing or not needing help to perform that ADL. For example, 1 question reads, "Dressing includes getting clothes from closets and drawers and putting the clothes on. Some people get dressed by themselves but may just need someone else to tie their shoes. Other people need someone to get the clothes, and to help them put them on . . . Please tell us if, currently, you think you need help with this activity most of the time."

After an explanation of standard-gamble utility ratings, subjects practiced rating 2 sample health states (having a head cold and being blind) and were offered an opportunity to discuss the procedure with the research assistant if they were unsure of the task. Subjects were then asked to give standard-gamble utilities for their current health (on a scale anchored by *perfect health* and *death*) and for 15 additional health states defined by 1 or more ADL dependency (on a scale from *cure* to *death*). All subjects were asked to rate the health states in the following order: each of the 7 single ADL dependencies, then the health state of dependence in all 7 ADLs, and finally 7 other combinations of ADL dependencies (2 combinations of 2 ADLs, 2 combinations of 3 ADLs, and 1 combination each of 4, 5, and 6 ADL dependencies). Within each group of single or multiple ADL dependencies (e.g., health states that consist of 2 ADL dependencies), the health states were presented to each subject in random order. Of the 127 potential combinations of the 7 ADL dependencies, we elicited utilities for the 30 combinations that account for 98% of ADL dependencies among the elderly (although each subject rated only 15 of those combinations).[31]

The risk of death that subjects were willing to accept to avoid dependence in a given ADL or combination of ADLs was obtained as follows. Subjects were asked if they would be willing to accept a 1% risk of death and 99% chance of cure. If they answered yes, they were asked the question again with probability values varied in a converging ping-pong fashion by alternating between high and low values: 90% risk of death/10% chance of cure, 10% risk of death/90% chance of cure, 80% risk of death/20% chance of cure, and so on.[32] If subjects answered no to the initial question of whether they would be willing to accept a 1% risk of death and 99% chance of cure, they were asked if there was any risk of death they would be willing to accept to avoid dependence in the health state.

The program required that subjects answer the first 2 questions with a yes or no response so that they would consider both a low risk of death and a high risk of death to minimize anchoring bias. After answering the first 2 yes/no (ping-pong) questions, subjects who had completed their response could directly provide their preferred risk by clicking on up/down arrows to change the numerical value presented to them. These utilities obtained on a scale with an upper anchor of *cure* could be rescaled to an upper anchor of *perfect health* using the current health utility; however, such rescaling is not necessary for an evaluation of consistency.

Subjects were given 2 opportunities to review their ratings and to make changes: first, after rating the health states of dependence in single ADLs and dependence in all 7 ADLs, and second, after rating the other combinations of ADL dependencies. The review screens showed all the ratings given in the previous segment and invited the subjects to review them and make any changes they wished. For example, subjects were asked, "In this section we will ask you to review all the ratings you have just done. The ratings are arranged from least to most chance of death that you said you would accept. The health condition 'All Activities' is the condition in which you need help with all 7 activities. We would expect that you would accept the highest risk for 'All Activities'. Think about each of the activities compared to the others. Would you like to change any of your ratings?"

Subjects were randomly selected to receive a computer-generated consistency reminder. If 1 of the 167 subjects (42%) selected to receive the reminder rated any health state worse than the combination of all 7 ADL dependencies, he or she received an alert showing his or her rating for dependence in all 7 ADLs and calling attention to the lower rating he or she had given to the other health state. He or she was then given an opportunity to alter his or her response. This reminder was presented only once for any inconsistent rating; that is, it did not force the subject to change his or her response. A sample size of 400 subjects yielded 80% power at an α level of 0.05 to demonstrate a 15% higher rate of consistency in the reminder group.

After the preference assessment, the research assistant administered a paper-based health status questionnaire (SF-36).[33] A research assistant was present for all portions of the interview. For the computer portion of the interview, research assistants answered technical questions about navigation with the trackball but did not help subjects interpret the utility questions (one of the reasons for using a computer-based elicitation was to standardize the elicitation process and to minimize interviewer bias).

Probability Transformation

All preferences were elicited as a probability of death (i.e., as a percent risk from 0% to 100% that the respondent was willing to accept). These values were converted to a utility on a scale from 0.0 (equivalent to death) to 1.0 (equivalent to perfect health for the current health rating and to cure of the ADL dependency for ADL health states) by computing (100 – the elicited probability)/100.

Invariance and Inconsistency Evaluations

Because only those utilities that vary from each other have the potential for inconsistent ordering, we first determined whether a subject rated every health state at a single utility level (invariant). We then assessed the remaining subjects' ratings for potentially inconsistent ordering. By "potentially inconsistent," we mean that we could not determine whether a subject may have had a valid reason for giving utilities that did not conform to our expected ordering. We defined "consistent" subjects as those who ordered utilities such that the utility for any single ADL dependency was greater than or equal to the utility for any combination of ADL dependencies in which it is contained (e.g., that the utility for dependence in bathing is greater than or equal to the utility for dependence in both bathing and continence). We also defined consistent subjects as those who ordered utilities such that the utility for being dependent in all 7 ADLs was less than or equal to any other utility given for any single ADL dependency or combination of dependencies. It is possible that an individual with an unusual preference structure might believe that having multiple ADL dependencies would be better than having just one because, for example, they believe that they would have more help available to them if they have multiple dependencies.

For each inconsistent response, we calculated the magnitude of inconsistency as the difference between the utility for the composite health state of multiple ADL dependencies and that of its inconsistent subset. Additionally, we calculated a maximum magnitude of inconsistency for each subject as the largest magnitude of inconsistency given by that subject.

Statistical Analyses

We calculated descriptive summary statistics (i.e., mean ± *s*) for subjects' demographics, current ADL dependencies, MMSE scores, time spent at the computer, and utility ratings. To evaluate the generalizability of our results to the Kaiser population, we compared the gender, race/ethnicity, and educational levels of the included subjects and the Kaiser Permanente population from which they were recruited using a *t* test. For this analysis, we used a Bonferroni correction to adjust for multiple comparisons such that the null hypothesis was rejected for a *P* value < 0.017 (0.05/3 = 0.017). We used multinomial logistic regression analysis to assess the likelihood of the 3 mutually exclusive outcomes: invariance, order inconsistency, and order consistency when not invariant, according to subject characteristics (i.e., age, gender, race/ethnicity, marital status, education, MMSE score, and number of current ADL dependencies); time spent during the computer elicitation; and whether subjects received the consistency reminder. We also used ordinary least squares regression to assess the association of these subject characteristics, whether the subjects received the consistency reminder, and time spent during the elicitation with the maximal magnitude of inconsistency.

RESULTS

Study Subjects

Of the 445 subjects who agreed to participate, 3 (1%) failed the MMSE, 9 (2%) refused to complete any of the computer study, 4 (1%) refused to complete part of the computer study, and 29 (7%) served as beta testers for the FLAIR2 elicitation tool. Our summary data are based on ratings from the remaining 400 subjects who completed the utility ratings for the single ADL health states using FLAIR1 (Tables 1 and 2). The average age of the included subjects was 73.2 y (*s* 5.7 y; range 65 to 91 y) and average MMSE score was 28.6 (*s* 1.6; range 24 to 30). Thirty-eight (10%) reported dependence in at least 1 ADL. When compared with the Kaiser population from which subjects were recruited (Table 1), the included subjects were more likely to be non-Hispanic White/Caucasian (*P* < 0.002) but were similar with respect to age and gender.

Patterns of Invariance

Seventy-six subjects (19%) gave a utility of 1.0 for all health states of ADL dependence. Nineteen subjects (5%) gave the same utility, other than 1.0, for all health states of ADL dependence (14 of whom gave a rating of 0.99 for all health states). These 95 subjects were all considered invariant and order consistent.

Table 1 Subject Demographics

	Included Subjects			Kaiser Population from Which Subjects Were Recruited[a]		
Characteristic	N^b	N_{total}	Percentage	N^b	N_{total}	Percentage
Female gender	222	400	56	583	1012	58
Race/Ethnicity[c]:						
Non-Hispanic White/Caucasian	367	397	92	771	1011	76
Other than non-Hispanic White/Caucasian	30	397	8	240	1011	24
Education (highest completed)						
8th grade or less	3	387	1	—	—	—
Some high school	94	387	24	—	—	—
High school graduates				637	1012	63
Some college	199	387	51	—	—	—
Some postgraduate education	91	387	24	—	—	—

a. The average age of the Kaiser population from which subjects were recruited was 74.4 y (*s* 6.8 y).
b. N = number of subjects for whom these results were recorded.
c. The included subjects were more likely to be non-Hispanic White/Caucasian than the eligible subjects who were not included ($P < 0.002$).

Order Inconsistency

One hundred thirty subjects (33%) gave some utilities less than 1.0 and had no order inconsistencies, whereas 175 subjects (44%) had at least 1 order inconsistency (Table 3). Ninety-one subjects (23%) provided a utility for dependence in all 7 ADLs that was greater than their utility for 1 or more of the single ADL dependencies (Table 3).

Magnitude of Inconsistency

The mean maximum magnitude of inconsistency was 0.19 (*s* 0.24; range 0.01 to 0.99). Among the inconsistent subjects, 15% had a maximal magnitude of inconsistency of 0.01, 54% had a maximal magnitude of inconsistency of 0.1 or less, and 69% had a maximal magnitude of inconsistency of 0.2 or less, which is equivalent to about 1 standard deviation about each of the mean utility ratings (Figure 1).

Inconsistency as a Function of Number of ADL Dependencies

To determine whether fatigue was primarily responsible for subjects' inconsistencies, we evaluated whether subjects provided their 1st inconsistent response early or late in the elicitation process. The 1st inconsistent utility for 91 subjects (23%) was for the combination of all 7 ADL dependencies, the 1st opportunity to demonstrate inconsistency. An additional 57 (14%) gave their 1st inconsistent response when they

Table 2 Mean Standard-Gamble Utilities for Current Health and ADL Dependencies

Health State	Mean	*s*
Current health[a]	0.89	0.21
Dependence in dressing	0.89	0.23
Dependence in bathing	0.88	0.24
Dependence in continence	0.86	0.25
Dependence in toileting	0.86	0.25
Dependence in transferring	0.86	0.26
Dependence in walking	0.85	0.25
Dependence in eating	0.85	0.27
Dependence in any single ADL dependency	0.86	0.25
Dependence in combinations of 2 ADL dependencies	0.83	0.27
Dependence in combinations of 3 ADL dependencies	0.82	0.27
Dependence in combinations of 4 ADL dependencies	0.81	0.28
Dependence in combinations of 5 ADL dependencies	0.79	0.29
Dependence in combinations of 6 ADL dependencies	0.79	0.30
Dependence in all 7 ADLs	0.76	0.32

Note: ADL = activity of daily living.
a. Whereas the rating for current health was on a scale from perfect health to death, the ratings for the other health states were on a scale from cure to death. The utilities obtained on a scale with an upper anchor of *cure* could be rescaled to an upper anchor of *perfect health* using the current health utility; however, such a rescaling is not necessary for consistency evaluation.

Table 3 Frequencies and Magnitudes of Order Inconsistencies

Type of Order Inconsistency	*n* (%)[a]	Mean Magnitude of Inconsistency (*s*)[b]
Utility for a single ADL dependence < Utility for dependence in 7 ADLs	91 (12)	0.18 (0.24)
Utility for a single ADL dependence < Utility for dependence in 2 ADLs	115 (16)	0.15 (0.21)
Utility for dependence in 2 ADLs < Utility for dependence in 7 ADLs	44 (6)	0.19 (0.28)
Utility for a single ADL dependence < Utility for dependence in 3 ADLs	94 (13)	0.17 (0.23)
Utility for dependence in 3 ADLs < Utility for dependence in 7 ADLs	41 (6)	0.18 (0.27)
Utility for a single ADL dependence < Utility for dependence in 4 ADLs	84 (11)	0.17 (0.22)
Utility for dependence in 4 ADLs < Utility for dependence in 7 ADLs	37 (5)	0.20 (0.28)
Utility for a single ADL dependence < Utility for dependence in 5 ADLs	85 (12)	0.17 (0.22)
Utility for dependence in 5 ADLs < Utility for dependence in 7 ADLs	42 (6)	0.19 (0.27)
Utility for a single ADL dependence < Utility for dependence in 6 ADLs	69 (5)	0.17 (0.24)
Utility for dependence in 6 ADLs < Utility for dependence in 7 ADLs	36 (5)	0.22 (0.29)

Note: ADLs = activities of daily living.

a. *n* refers to the number of order-inconsistent responses. The denominator for the percentage calculation is 738, referring to the total number of inconsistencies.

b. Mean (*s*) magnitude of inconsistency for each type of order inconsistency.

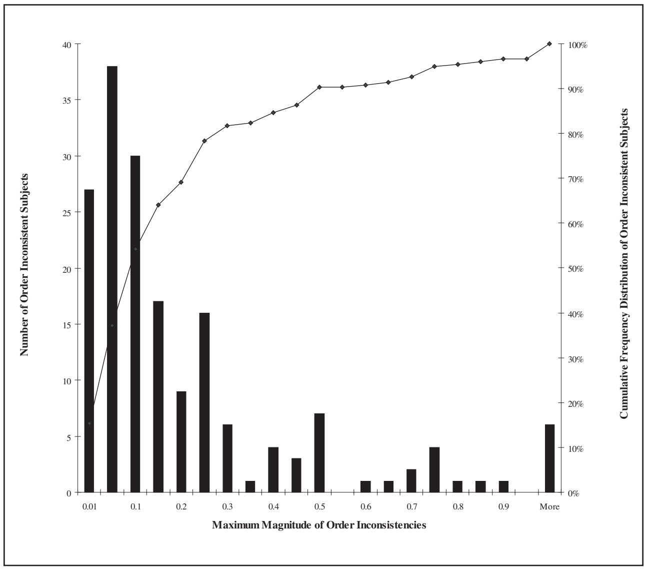

Figure 1 *Maximal magnitude of order inconsistency. The maximum magnitude of inconsistency is calculated as the maximum difference between the utility for the combined activity of daily living (ADL) dependence health state and that of its inconsistent subset. In this figure, the maximal magnitude of order inconsistency per subject along the x-axis is plotted against the number of order-inconsistent subjects on the left-hand y-axis (denoted by the black bars), and the cumulative frequency distribution of order-inconsistent subjects is plotted on the right-hand y-axis (denoted by the small squares and the line).*

assigned a utility to a combination of 2 ADL dependencies, the 2nd opportunity for inconsistency. Thus, 148 of the 175 inconsistent subjects (85%) had already provided inconsistent ratings before beginning to rate health states of 3 or more ADL dependencies.

Predictors of Invariance and Inconsistency

The average time subjects spent with the computer elicitation program was 40.7 min (*s* 15.8 min). On average, the invariant subjects spent 31.6 min (*s* 14.6 min) on the computer elicitation program, those who were order consistent but not invariant spent 42.0 min (*s* 12.9 min), and those subjects who were order inconsistent spent 44.8 min (*s* 16.6 min; $P < 0.0001$ for difference among the 3 groups).

In a multinomial logistic regression with the dependent variable comprising inconsistency compared with consistency and invariance compared with consistency, the consistency reminder was not associated with greater consistency and no subject characteristic was associated with order inconsistency (Table 4). Invariance was associated with below-average MMSE scores (less than 28.6), with lack of current dependence in any ADLs, with below-average education levels (less than 12 y), with race/ethnicity other than non-Hispanic White/Caucasian, and with a relatively short time spent on the computer elicitation program. In a linear regression with maximal magnitude of inconsistency as the dependent variable, only the coefficients for time spent during the computer elicitation and race/ethnicity other than non-Hispanic White/Caucasian were significant ($P < 0.0001$ and 0.005, respectively; adjusted $R^2 = 0.19$).

Differences in Utilities between Consistent and Inconsistent Subjects

The consistent subjects tended to have higher utility ratings for all states presented to them than the incon-

Table 4 Odds of Being Invariant Compared with Being Consistent and of Being Order Inconsistent Compared with Being Consistent

	Odds of Being Invariant Compared with Being Consistent		Odds of Being Order Inconsistent Compared with Being Consistent	
Factors	**OR**	**95% CI**	**OR**	**95% CI**
Received consistency reminder	1.31	(0.70, 2.44)	1.14	(0.70, 1.84)
Age greater than or equal to 73.2 y[a]	0.62	(0.33, 1.16)	0.73	(0.45, 1.18)
Mini-Mental State Examination score greater than 28.6[a]	0.45	(0.24, 0.82)	1.01	(0.61, 1.66)
Currently dependent in at least 1 ADL	0.33	(0.11, 0.98)	0.85	(0.38, 1.90)
Greater than 12 y of education	0.38	(0.20, 0.73)	0.75	(0.43, 1.30)
Female gender	1.01	(0.60, 1.70)	0.96	(0.63, 1.47)
Race/ethnicity other than non-Hispanic White/Caucasian	4.60	(1.85, 11.45)	1.82	(0.82, 4.03)
Time spent on the computer elicitation greater than average	0.09	(0.04, 0.20)	1.16	(0.71, 1.90)

Note: Goodness of fit for this model: χ^2 108.0, 16 *df*, $P < 0.0001$; OR = odds ratio; ADL = activity of daily living.
a. For age and Mini-Mental State Examination scores, these values represent the mean for all 400 subjects.

sistent subjects. The mean (*s*) utility for dependence in continence among consistent subjects who were not invariant (0.88 [0.24]) was significantly higher than among inconsistent subjects (0.80 [0.27]; $P = 0.01$).

DISCUSSION

This 400-subject study, which assessed utility ratings for 15 health states that have an implicit ordering, offers an unusual opportunity to examine invariance and inconsistency in assessed preferences. Our results contribute to the literature in 3 ways: First, we comprehensively evaluated utilities for invariance and order inconsistencies. Second, we sought both patient characteristics and factors associated with the elicitation task itself (such as the reminder and the time spent on the task) that predicted invariance and inconsistency. Understanding these factors can inform efforts to improve elicitation tasks and minimize invariance and inconsistency. Finally, we evaluated the extent to which the utilities of inconsistent, invariant, and consistent (but noninvariant) subjects differed. This information is important when considering whether to report utilities from all subjects or just from a sample of subjects (e.g., the consistent subjects).

Our study produced 5 key findings: First, our primary hypothesis was confirmed—even among a large number of relatively well-educated subjects, invariance and order inconsistencies are common. Our finding that 44% of the subjects had at least 1 order inconsistency is within the range of published utility elicitations that have performed similar inconsistency assessments (12% to 59%).[1,8–11,15] The finding that 5% of our subjects gave an invariant, non-1.0 rating for all health states is also similar to a previous study.[8] Second, subjects who give order-inconsistent utility ratings do so early in the elicitation process and are likely to take longer at the elicitation task than subjects without order inconsistencies. Third, a simple consistency reminder does not improve order consistency. Fourth, the maximum magnitude of order inconsistencies for most subjects is relatively small. Finally, the utilities that consistent subjects assign to certain health states may systematically differ from the ratings of inconsistent subjects, raising questions about whether to include inconsistent ratings. We note that in our dataset the ratings of order-inconsistent subjects fell within 1 standard deviation of mean utilities for all subjects, so that including the ratings of the inconsistent subjects in the calculation of the mean utility ratings would not change the overall mean utility ratings by much.

Invariance

Nearly 20% of our subjects assigned a utility of 1.0 to all health states. We note that despite the finding that the invariant subjects spent less time at the elicitation task than other subjects, the 76 subjects giving fixed utility ratings of 1.0 were not providing the “response of least resistance” because they had to click through several questions to get to a response indicating 1.0. This type of invariance—with all utility ratings equal to 1.0—is rarely reported in the literature possibly because multiple utilities must be elicited from the same individual to demonstrate invariance. There are, however, previous reports of utility ratings that include a surprisingly high proportion of ratings equal to 1.0. For example, Tsevat and colleagues found that 35% of 1348 seriously ill hospitalized patients who provided time-tradeoff utilities for current health had utilities of 1.0.[34] Thus, we believe that this type of invariance may be more common in utility assessments than previously recognized. We recommend that utility elicitations with multiple health states include assessments of invariance. If an invariant utility of 1.0 is more common than previously recognized, it may account for some of the high utilities reported for even severe health states (particularly among standard-gamble assessments).[35–37]

All subjects who completed the computer ratings appeared to the research assistants to understand the rating procedure, thereby meeting the usual criterion for inclusion in utility elicitations. However, it is possible that at least some of these subjects did not understand the full complexity of the task. We found that invariance was associated with lower MMSE scores, education levels, and race/ethnicity other than non-Hispanic White/Caucasian—factors similar to those that have been associated with poor task comprehension and higher rates of inconsistencies in other preference assessments. This highlights the need for evaluating whether such subjects understand the rating task and are truly unwilling to accept any risk of death (invariant, utility = 1.0) or have a set amount of risk that they are willing to accept for most health states (invariant, utility < 1.0). Invariant respondents could be asked to provide reasons for their responses. In addition, utilities for extreme undesirable health states could be elicited for comparison with the target health state: some respondents may be willing to risk dying for such states but not for milder states, suggesting that their rating of utility = 1.0 for the target health state is valid. Invariant subjects could be asked to provide additional preferences for the target health states by other methods (e.g., willingness-to-pay) to determine whether they would provide varying ratings if a different measure of overall value were used in the case of invariance with a utility of 1.0, or if there is any health state that is so seriously impaired that they would assign a rating lower than 1.0.

Order Inconsistency

Although order inconsistencies were common among our subjects, our findings about the magnitudes of inconsistency demonstrate that, for many inconsistent subjects, the maximum magnitude of inconsistencies was relatively small. Given the large number of health states that subjects were asked to rate, even with the opportunity to review their ratings, it may not be surprising that subjects were found to have small inconsistencies.

Including the ratings of inconsistent subjects with small magnitudes of inconsistency may have only modest effects on mean reported utilities. Because there is no consensus about including utilities of inconsistent subjects in calculating mean utilities, it is prudent to report the magnitude of inconsistency and to determine whether mean utilities differ between consistent and inconsistent subjects to allow users of utility data to choose whether to include or exclude subjects with order inconsistencies. It is unclear why the consistent and inconsistent subjects should differ in their utilities for dependence in continence. We regard this as a preliminary finding that requires confirmation.

We found that the rate of order inconsistency was essentially the same between the subjects who received the consistency reminder and those who did not, a nonsurprising finding given the weak nature of the reminder, which did not force order consistency. We are currently evaluating whether reminding subjects of their ranking of health states on a visual analog scale is more likely to result in greater order consistency.

When we observed that the order-inconsistent subjects spent more time at the elicitation task than consistent subjects, we wondered whether order inconsistencies may have resulted from subject fatigue. Subjects in our study rated a relatively large number of health states. However, fatigue alone is unlikely to be responsible in that 85% of order-inconsistent subjects provided an inconsistent response by the time they had rated health states of 2 ADL dependencies, before they had reached the more taxing tasks of weighing health states with particular combinations of 3, 4, or 5 ADL dependencies.

Although invariance and inconsistencies could in part be a consequence of the assessment instrument, the FLAIR1 program was designed to minimize such effects. FLAIR1 was tested extensively on older adults to improve usability and was designed with multiple consistency-enhancing measures. First, it is a multimedia tool incorporating sound, photographs, animated graphics, video, and text that provides a detailed description of hypothetical health states, allowing the respondent to form a clear image of the health state while preserving uniformity from 1 presentation to the next. Second, the program included priming activities to familiarize the respondents with the health states and encourage them to think seriously about the health states by providing responses about their own dependence or independence in the activities. Third, to minimize framing effects, all risks were shown in both positive and negative terms, that is, as both risk of death and chance of perfect health/cure. Fourth, to minimize anchoring effects, the respondents were required to consider yes/no questions presenting both high and low risks of death. Fifth, the program allowed respondents to review previously rated health states. Finally, the program was designed to be easy for subjects with no computer experience and to be accessible for subjects with arthritis or poor vision: respondents were required only to move a large trackball and depress a single button.

Our study had 2 primary limitations. First, 48% of the randomly selected subjects chose not to participate. Although the included subjects were similar to the Kaiser population from which they were recruited with respect to age and gender, the included subjects were more likely to be non-Hispanic White/Caucasian. Thus, our results may not be generalizable to the entire spectrum of Kaiser Permanente patients or, indeed, to all elderly patients. Second, we do not have interview or other data to determine whether the utility ratings are consistent with the beliefs and preferences of the subjects. To address both of these limitations, in our ongoing evaluation of FLAIR2, we have oversampled subjects with race/ethnicity other than non-Hispanic White/Caucasian and are performing exit interviews.

Our data set allowed us to explore trends in invariance and order inconsistency on a larger scale than previous preference assessments. The high prevalence of invariance and inconsistency is not unique to these data. Consequently, we recommend that preference assessments measure and report both invariance and inconsistency. If including invariant and inconsistent responses does not alter mean reported utilities, researchers may choose to include these ratings. However, if subjects have large magnitudes of inconsistencies or if inconsistent responses differ significantly from mean responses, we recommend that researchers report mean utilities both with and without these subjects.

ACKNOWLEDGMENTS

We thank Pamela Mahlow, MA, Sheryl Davies, MS, and Manijeh Parineh, MPH, for their assistance with the subjects, data collection, and data preparation; Deborah McNaughton for her statistical review of this article; and David Miller, PhD, for his contribution to the development of the FLAIR1 program. We also thank Joseph Selby, MD, MPH, and the FLAIR staff at the Division of Research, Kaiser Permanente Medical Care Program, for assistance in identifying individuals to participate in this study. The views expressed in this article are those of the authors, not the Department of Veterans Affairs or other affiliated or participating institutions. Dr Goldstein's work was supported by a Career Development Award from the VA Health Services Research and Development Service HSR&D ARCD-96301. The FLAIR project is supported by National Institutes for Health grant AG15110.

REFERENCES

1. Badia X, Roset M, Herdman M. Inconsistent responses in three preference-elicitation methods for health states. Soc Sci Med. 1999;49:943–50.
2. Lenert LA, Cher DJ, Goldstein MK, Bergen MR, Garber A. The effect of search procedures on utility elicitations. Med Decis Making. 1998;18:76–83.
3. Lenert LA, Morss S, Goldstein MK, Bergen MR, Faustman WO, Garber AM. Measurement of the validity of utility elicitations performed by computerized interview. Med Care. 1997;35:915–20.
4. Froberg DG, Kane RL. Methodology for measuring health-state preferences—IV: progress and a research agenda. J Clin Epidemiol. 1989;42:675–85.
5. Llewellyn-Thomas HA, McGreal MJ, Thiel EC. Cancer patients' decision making and trial-entry preferences: the effects of "framing" information about short-term toxicity and long-term survival. Med Decis Making. 1995;15:4–12.
6. Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC. Preferences for health outcomes: comparison of assessment methods. Med Decis Making. 1984;4:315–29.
7. Eraker SA, Sox HC Jr. Assessment of patients' preferences for therapeutic outcomes. Med Decis Making. 1981;1:29–39.
8. Rutten-van Molken MP, Bakker CH, van Doorslaer EK, van der Linden S. Methodological issues of patient utility measurement. Experience from two clinical trials. Med Care. 1995;33:922–37.
9. Lenert LA, Sturley A, Rupnow M. Toward improved methods for measurement of utility: automated repair of errors in elicitations. Med Decis Making. 2003;23:67–75.
10. Giesler RB, Ashton CM, Brody B, et al. Assessing the performance of utility techniques in the absence of a gold standard. Med Care. 1999;37:580–8.
11. Souchek J, Stacks JR, Brody B, et al. A trial for comparing methods for eliciting treatment preferences from men with advanced prostate cancer: results from the initial visit. Med Care. 2000;38:1040–50.
12. Gafni A. The standard gamble method: what is being measured and how it is interpreted. Health Serv Res. 1994;29:207–24.
13. Nickerson CA. Assessing convergent validity of health-state utilities obtained using different scaling methods. Med Decis Making. 1999;19:487–98.
14. Lenert LA, Treadwell JR. Effects on preferences of violations of procedural invariance. Med Decis Making. 1999;19:473–81.
15. Goossens ME, Vlaeyen JW, Rutten-van Molken MP, van der Linden SM. Patient utilities in chronic musculoskeletal pain: how useful is the standard gamble method? Pain. 1999;80:365–75.
16. Lee TT, Ziegler JK, Sommi R, Sugar C, Mahmoud R, Lenert LA. Comparison of preferences for health outcomes in schizophrenia among stakeholder groups. J Psychiatr Res. 2000;34:201–10.
17. Lenert LA, Ziegler J, Lee T, Sommi R, Mahmoud R. Differences in health values among patients, family members, and providers for outcomes in schizophrenia. Med Care. 2000;38:1011–21.

18. Woloshin S, Schwartz LM, Moncur M, Gabriel S, Tosteson AN. Assessing values for health: numeracy matters. Med Decis Making. 2001;21:382–90.

19. Dolan P, Kind P. Inconsistency and health state valuations. Soc Sci Med. 1996;42:609–15.

20. Keeney RL, Raiffa H. Decisions with Multiple Objectives: Preferences and Value Tradeoffs. Cambridge (UK): Cambridge University Press; 1993.

21. Patrick DL, Sittampalam Y, Somerville SM, Carter WB, Bergner M. A cross-cultural comparison of health status values. Am J Public Health. 1985;75:1402–7.

22. Sanders GD, Owens DK, Padian N, Cardinalli AB, Sullivan AN, Nease RF. A computer-based interview to identify HIV risk behaviors and to assess patient preferences for HIV-related health states. Proc Annu Symp Comput Appl Med Care. 1994:20–4.

23. Wang Q, Furlong W, Feeny D, Torrance G, Barr R. How robust is the Health Utilities Index Mark 2 utility function? Med Decis Making. 2002;22:350–8.

24. Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. Med Care. 1996;34:702–22.

25. Stalmeier PF, Goldstein MK, Holmes AM, et al. What should be reported in a methods section on utility assessment? Med Decis Making. 2001;21:200–7.

26. Katz S, Ford A, Moskowitz R, Jackson B, Jaffe M. Studies of illness in the aged: the index of ADL: a standardized measure of biological and psychosocial function. JAMA. 1963;185:94–9.

27. McDowell I, Newell C. Measuring Health: A Guide to Rating Scales and Questionnaires. Vol. 1, 2nd ed. New York: Oxford University Press; 1996.

28. Goldstein MK, Miller DE, Davies S, Garber AM. Quality of life assessment software for computer-inexperienced older adults: multimedia utility elicitation for activities of daily living. Proc AMIA Symp. 2002:295–9.

29. Goldstein MK, Tsevat J. Applying utility assessment at the "bedside". In: Chapman G, Sonnenberg FA, eds. Decision Making in Health Care: Theory, Psychology, and Applications. Cambridge (UK): Cambridge University Press; 2000.

30. Goldstein MK, Tsevat J. Assessing desirability of outcome states for medical decision making and cost-effectiveness analysis. In: Max M, Lynn J, eds. Symptom Research: Methods and Opportunities. Available from http://symptomresearch.nih.gov. National Institutes of Health; 2003.

31. NLTCS. Limitations in Activities of Daily Living among the Elderly: Data Analyses from the 1989 National Long-Term Care Survey (HHS-100-95-0017). Washington (DC): Administration on Aging, US Department of Health and Human Services; May 31, 1996.

32. Dolan P, Gudex C, Kind P, Williams A. Valuing health states: a comparison of methods. J Health Econ. 1996;15:209–31.

33. Ware JE Jr, Sherbourne CD, Stewart AL, Hays RD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992;30:473–83.

34. Tsevat J, Cook EF, Green ML, et al. Health values of the seriously ill. SUPPORT investigators. Ann Intern Med. 1995;122:514–20.

35. Tengs TO, Lin TH. A meta-analysis of quality-of-life estimates for stroke. Pharmacoeconomics. 2003;21:191–200.

36. Tengs TO, Lin TH. A meta-analysis of utility estimates for HIV/AIDS. Med Decis Making. 2002;22:475–81.

37. Tengs TO, Wallace A. One thousand health-related quality-of-life estimates. Med Care. 2000;38:583–637.