ELSEVIER

Journal of Health Economics 15 (1996) 209–231

JOURNAL OF HEALTH ECONOMICS

Valuing health states: A comparison of methods

P. Dolan [a,*], C. Gudex [b], P. Kind [b], A. Williams [b]

[a] *Department of Economics, University of Newcastle, Newcastle-Upon Tyne, UK*
[b] *Centre for Health Economics, University of York, Heslington, York, UK*

Received 15 August 1993; revised 15 June 1995

Abstract

In eliciting health state valuations, two widely used methods are the standard gamble (SG) and the time trade-off (TTO). Both methods make assumptions about individual preferences that are too restrictive to allow them to act as perfect proxies for utility. Therefore, a choice between them might instead be made on empirical grounds. This paper reports on a study which compared a "props" (using specially-designed boards) and a "no props" (using self-completion booklets) variant of each method. The results suggested that both no props variants might be susceptible to framing effects and that TTO props outperformed SG props.

JEL classification: I1

Keywords: Health state valuation; Time trade-off; Standard gamble; Validity; Reliability

1. Introduction

Given that no country can afford to provide all the health care that might conceivably be of some benefit, it is necessary to establish priorities. Although there is no consensus as to how this priority-setting should be done, there is general agreement that the impact on the health status of the population affected by alternative uses of scarce resources should be taken into account. Against this background, a form of economic appraisal, referred to as cost-utility analysis

* Corresponding author. Fax: +44 191 222-6548.

SSDI 0167-6296(95)00038-0

(CUA), has been developed to compare the costs of a health care programme with its benefits, measured in terms of its impact on both length of life *and* quality of life.

There are two main stages in the development of any measure of health-related quality-of-life (HRQoL) that is used in CUA. The first is to describe health status in terms of domains, or dimensions. There is some consensus now regarding the dimensions that are relevant to measures aimed at describing general health status (Bullinger, 1991; Berzon and Shumaker, 1993) The second stage involves determining the numerical value to be attached to the health states described. CUA requires that cardinal values be assigned to each health state on a scale where "full health" is assigned a value of 1 and "death" a value of 0. To allow for the possibility that some states may be regarded as being worse than death, negative values should also be allowed for.

In determining the values attached to different health states, analysts can adopt one of three strategies: (1) use expert judgement, (2) use values obtained from relevant literature, or (3) use direct measurement (Torrance, 1986). Because of the potential sources of bias associated with the first two (for example, judgements may be wrong, or published literature may be inappropriate) the third strategy is generally seen as the most appropriate. However, there exist a number of different procedures for eliciting valuations and there appears to be little agreement as to which is the preferred technique. The choice of method is important, as different methods appear to yield different sets of valuations for identical descriptions of quality of life (Torrance, 1976; Wolfson et al., 1982; Read et al., 1984; Hornberger et al., 1992). This paper focuses on an empirical study designed to test the performance of different valuation methods, and to determine which method should be chosen for wider use.

Of the many methods available for measuring health state utilities, two of the most widely used have been the standard gamble (SG) and the time trade-off (TTO). Both methods start from the premise that, given that health is an important argument in an individual's utility function, we can estimate the welfare change associated with a change in health if we can determine the compensating change in one of the remaining arguments in an individual's utility function that leaves utility unchanged. In the SG, health improvements are valued in terms of the level of risk (usually of immediate death) an individual is prepared to accept, which means assuming utility to be a negative function of such a risk. In the TTO, health improvements are valued in terms of the amount of life expectancy an individual is prepared to sacrifice by assuming utility to be a positive function of longevity. In this way, both the SG and the TTO can be viewed as sharing a common theoretical background.

That the SG and TTO do yield different valuations from the same respondents is evidenced by most empirical studies to date. Torrance (1976) and Read et al. (1984) found correlations of 0.65 between the scores elicited by the two methods. Torrance concluded that the two methods are equivalent, but Read et al. empha-

sised that high correlations can coexist with systematic differences between sets of scale values. Wolfson et al. (1982) found that SG utilities were consistently higher than TTO values whilst, in a more recent study, Hornberger et al. (1992) found much poorer correlation between SG and TTO, particularly at the individual level. Froberg and Kane (1989b) concluded that "while correlations between methods are usually moderately high, the different methods do not necessarily produce equivalent scale values".

In developing Expected Utility Theory (EUT), von Neumann and Morgenstern (1953) showed that if a cardinal utility could be expressed as equivalent to a gamble, under certain assumptions, it would be a linear function of the risk involved in the gamble. In other words, the level of risk involved in standard gamble questions is linear in utility. This led many to regard the SG as the "gold standard" for health status measurement (Torrance, 1976). However, doubt has been cast on EUT both as a positive *and* as a normative theory. First, there is evidence that people systematically violate the axioms of EUT (Llewellyn-Thomas et al., 1982; Schoemaker, 1982). Thus, much of the appeal of the SG is lost since it will only be an accurate measure of utility *if* the axioms of EUT apply. Second, EUT focuses *only* on the expected utility of different outcomes, and there is increasing evidence that many people consider this to be an *irrational* basis on which to make decisions, preferring instead to take account of the process by which the outcomes were arrived at (Loomes and Sugden, 1982).

The literature often distinguishes between utility, which results from decisions under uncertainty (as measured by the SG, for example), and value, which results from decisions based on certainty (Gafni et al., 1993). Because in the TTO method both of the alternatives presented to the respondents have outcomes that are known with certainty, it is said to produce a value, not a utility, function (Pliskin et al., 1979; Bennet et al., 1991). However, this is based on a very narrow definition of utility, one that has arisen as a direct result of Von Neuman Morgenstern EUT. In its broader sense, and one which is perhaps more relevant to the measurement of quality-of-life, utility is defined as a (cardinal) index of strength of preference. It is possible to measure this under conditions of uncertainty *or* certainty.

The SG is also advocated on the grounds that almost all decisions about health care are made under conditions of uncertainty (Mehrez and Gafni, 1991). Whilst this is indeed the case, the appropriateness or otherwise of a valuation method is determined by its ability to act as a proxy for utility and not by its capacity to model the situation being valued (Buckingham and Drummond, 1993). In this respect, the TTO may be considered more appropriate since, by definition, it gives the number of years in full health which are valued equally to a (longer) period in the health state being measured. Thus, it collapses the relationship between the health state, its duration and its value into one single measure. Nevertheless, there is doubt about the validity of the underlying assumption of the TTO method that individuals are prepared to trade-off a constant proportion of their remaining years

of life in order to improve their health status, irrespective of the number of years that remain (Sackett and Torrance, 1978; Sutherland et al., 1982).

It is therefore difficult to choose between SG and TTO on theoretical grounds since valuations from neither method can automatically be assumed to map directly onto utility. This is an important point since it implies rejecting the idea that the SG should be regarded as the "gold standard" for measuring health state values. Instead, a choice between the SG and the TTO needs to be informed by their respective performance on empirical grounds. The evidence here is limited since relatively few studies have obtained within-respondent comparisons of the different valuation methods (Torrance, 1976; Wolfson et al., 1982; Read et al., 1984; Hornberger et al., 1992). Empirical assessment of the different techniques involves considerations of feasibility, consistency, validity and reliability.

Feasibility means that the method must be capable of being carried out in practice and be acceptable to respondents. This last point would appear to be satisfied by the high response rates and even higher levels of complete data that most studies have reported (Froberg and Kane, 1989a). *Consistency* refers to the extent to which the health states used in a study are given a logical ordering within a method. This might be seen as construct validation in the sense that it tests the construct that "better" states of health should be given higher scores but since this has rarely been considered (in fact, inconsistent respondents have generally been excluded from data analysis (Martin and Elliot, 1992; Torrance et al., 1992)) it is treated here as a criterion in its own right.

Essentially a measure is *valid* if it accurately reflects the concept or phenomenon it claims to measure. In establishing the validity of different methods, most studies have examined the extent to which the different methods yield similar results. This test, often referred to in the literature as concurrent validity, has been predicated on the notion that the SG represents the gold standard against which different methods are compared. Indeed, Torrance (1976) advocated the use of the TTO primarily *because* he found it to be correlated with the SG. The above discussion argues that the theoretical justification for according the SG such status is questionable. In this context, concurrent validity is an almost meaningless concept since it tells us nothing about which method is more valid if the methods yield different results, nor whether both or neither method is valid if the methods yield similar results. However, if one method yields very different results from a number of other methods, then doubt *may* be cast on its validity.

In the absence of a gold standard, the most rigorous approach to establishing validity is testing *construct validity*. A construct is a theoretically derived notion of what the method is intended to measure. An understanding of the construct allows the extent to which the method fulfils its predictions to be examined. Construct validity can be assessed by examining (a) the extent to which the valuations from the different methods are correlated with factors for which there is an a priori expectation of good correlation (sometimes referred to as *convergent validity*) and (b) the extent to which the valuations are *not* correlated with factors

for which there is expected to be poor correlation (sometimes referred to as discriminant validity).

The evidence currently available suggests that variation among population subgroups is not explained by the different demographic characteristics of respondents, such as age, sex, or socio-economic status (Carter et al., 1976; Kaplan et al., 1978; Rosser and Kind, 1978). There is, however, some evidence to suggest that experience of illness may influence respondents' valuations of health states. For example, Sackett and Torrance (1978) reported that home dialysis patients assigned higher utility to kidney dialysis than did the general public. In addition, Rosser and Kind (1978), from comparisons of patients, nurses, physicians and the general public found significant differences between medical patients and physicians and between medical patients and psychiatric patients. The possibility that valuations differ according to illness experience has been noted by Froberg and Kane (1989a) who state that "We have seen that patients with a particular condition often assign a higher utility than do patients without the condition".

The *reliability* of a valuation method can be investigated in two ways; (a) *Split-test reliability* which assesses an individual respondent's consistency when an item is presented more than once and (b) *Test-retest reliability* which assesses the stability of values over short periods of time. Torrance (1976) found the SG and TTO to have similar split-test correlation coefficients (between 0.80 and 0.90) and these results have been considered to be "acceptable" (Froberg and Kane, 1989b). O'Connor et al. (1985) reported correlations of 0.80 to 0.87 for a one week retest of SG and TTO, respectively, although some respondents may have remembered their initial valuations given the relatively short time interval between test and retest.

In view of the inconclusive results from the literature, the aim of this study was to identify the "best" *one* method for valuation of generic health states by interview. The SG and TTO were each tested in two variants, one of which used specially designed boards and cards as an aid to decision-making by respondents (props), and the other used a self-completed booklet (no props). Thus, four main methods were compared in this study. Against this background, four principal criteria were selected as the basis for choosing between competing valuation methods. These concerned the quality of the data elicited from the respondents, rather than the practical aspects of administering the different tasks (such as the burden placed upon respondents and interviewers) and were as follows:

1. *Completeness* (as a measure of feasibility): the extent to which each method produces a complete data set.
2. *Logical Consistency*: the extent to which the health states used were given a logical ordering within each method.
3. *Construct Validity*: the extent to which valuations differ in accordance with prior expectations.
4. *Test–retest Reliability*: the extent to which respondents' responses are stable within each method over a relatively short time interval.

2. Method

2.1. Study population

The sample was drawn from adults aged 18 and over in the general population. Anticipating a response rate of around 50%, a random sample of 700 addresses was drawn from 11 regional areas in the UK using the Postcode Address File. Twenty-five specially trained interviewers visited these homes and requested one interview from a randomly selected adult at each address. Of the 700 addresses issued to interviewers, 88 were non-residential or empty and were thus excluded. Of the remainder, 87 (14%) yielded no contact or were otherwise unproductive, 190 (31%) yielded refusals, and 335 (55%) yielded an interview. A sub-sample of those respondents to the study who had said they would be willing to be re-interviewed were approached again 6 to 16 weeks after the original interview. Respondents were asked to do exactly the same tasks as before, with the additional question of whether anything important had happened to them since the last interview. Of these, 20 (13%) yielded no contact or were otherwise unproductive, 26 (17%) yielded refusals, and 110 (71%) yielded an interview.

2.2. The interview

In order to try out the two variants of the SG and TTO, and to test for possible order effects, each interviewer was randomly allocated to one of eight experimental groups. All interviews used the Euroqol descriptive system, which describes health status in five dimensions with no disease specificity (see Table 1). The health states are described by combining one statement from each of these dimensions. The main aim of the Euroqol Group is to develop a generic measure that provides a simple but standard description of health status with a set of valuations that have interval properties (Euroqol Group, 1990). As part of its development, the Euroqol is being used to assess health status in a wide range of patient groups as well as in the general public, and alongside other generic and disease-specific health status measures (Essink-Bot, 1990; Munro et al., 1992; Brazier et al., 1993; Sculpher et al., 1993).

As a preliminary task, respondents described their own state of health in terms of the Euroqol classification (see Table 1) and then ranked eight health states (11111, 21111, 11122, 21221, 21232, 22323, 33333 and Immediate Death) from best to worst. Respondents then rated their own health "today" on a visual analogue scale (VAS), with 100 (best imaginable health state) and 0 (worst imaginable health state) as endpoints, and were then asked to place the eight states on the same scale such that the intervals between the placements corresponded to the differences perceived by the respondent. The principal purpose of the VAS was to familiarise respondents with the health state descriptions and to check on the comparability of the different experimental groups. The VAS was not one of

Table 1
The Euroqol descriptive system

Code number	Description
Mobility	
1.	No problems walking about
2.	Some problems walking about
3.	Confined to bed
Self-care	
1.	No problems with self-care
2.	Some problems washing or dressing self
3.	Unable to wash or dress self
Usual activities	
1.	No problems with performing usual activities (e.g. work, study, housework, family or leisure activities)
2.	Some problems with performing usual activities
3.	Unable to perform usual activities
Pain / discomfort	
1.	No pain or discomfort
2.	Moderate pain or discomfort
3.	Extreme pain or discomfort
Anxiety / depression	
1.	Not anxious or depressed
2.	Moderately anxious or depressed
3.	Extremely anxious or depressed

For convenience each composite health state has a five digit code number relating to the relevant level of each dimension, with the dimensions always listed in the order given above. Thus 11223 means:
1 No problems walking about;
1 No problems with self-care;
2 Some problems with performing usual activities;
2 Moderate pain or discomfort;
3 Extremely anxious or depressed.

the contenders in the choice of "best" valuation method because valuations from this technique are elicited in a choice*less* context. Thus, they do not reflect the importance of health relative to other arguments in an individual's utility function and are not regarded as measures of utility, defined in its broadest sense.

In order to achieve within-respondent comparisons, each respondent then did both a SG *and* a TTO exercise on six health states (the same states as in the ranking and scaling exercises, but excluding the 11111 state and Immediate Death). These six states were always presented in a fixed standard order. Finally respondents completed socio-demographic questions, and both respondents and interviewers gave feedback on the overall interview experience. Throughout the interview respondents were asked to imagine that each state would last for 10 years without any change and then they would die.

The SG asks the respondent to choose between the certainty of an intermediate health state and the uncertainty of a treatment with two possible outcomes, one of which is *better* than the certain outcome and one of which is *worse*. The object is to find the probability at which the respondent is indifferent between accepting the intermediate health state and accepting the treatment. When the two possible outcomes of the treatment are full health and death they are arbitrarily assigned utilities of 1 and 0, respectively. The utility of an intermediate state, h_b, that is rated as better than death is then determined by the probability at which the respondent is indifferent between h_b and the treatment. The (negative) utility for a state, h_w, rated as worse than death is determined by the probability mix of full health and h_w that makes the respondent indifferent between this and immediate death.

The *SG Props* (SGP) variant used here was based on the McMaster Chance Board (Furlong et al., 1989) but after pre-piloting was modified so that a sliding scale (instead of a chance wheel) showed the varying chances of success and failure of treatment. The same side of the board was used for states better and worse than death. After an explanation of the method using an example card, the interviewer presented each of the 6 states in their standard order. The chances of success of treatment were presented in a "ping-pong" fashion i.e. 100% success, 10% success, 90% success etc. The *SG No Props* (SGNP) variant consisted of a self-completed booklet for each state in which the respondent identified the chance of success at which (s)he would choose the treatment rather than stay in the health state. The respondent was taken through an example by the interviewer and then given the booklets in the standard order.

The TTO (as developed by Torrance, 1976) asks the respondent to choose between two alternatives both of whose outcomes are known with certainty. For a state, h_b, that is rated as better than death, the first alternative is to live for a defined period of time (for the purposes of this study, 10 years) in h_b and then die. The second alternative is to live for a shorter period of time (i.e. up to 10 years) in full health and then die. The time in full health is varied until the respondent is indifferent between the longer but lower quality life and the shorter but higher quality life. For a state, h_w, that is rated as worse than death, the first alternative is to die immediately and the second alternative is x years in h_w followed by y years in full health (where in this case x and y sum to 10 years). This differs from Torrance (1986) who gives an example where full health is experienced *before* the state rated as worse than death.

For *TTO Props* (TTOP) a sliding-scale on a board showed the number of years in good health compared to 10 years in the poor health state. One side of the board was used for states better than death and the other for states worse than death. After an explanation of the method using an example card, the interviewer presented each of the 6 states in their standard order. The number of years lost were also presented in a "ping-pong" fashion i.e. 0 years lost, 10 years lost, 1 year lost etc. For *TTO No Props* (TTONP) a self-completed booklet was used for

the state in which the respondent indicated the number of years of life that (s)he would be prepared to sacrifice in order to be in full health. The respondent was taken through an example by the interviewer and then given the booklets in the standard order.

3. Results

3.1. Characteristics of respondents

Table 2 shows that, although there were no significant differences between the test and retest respondents with respect to background or health variables, there was some response bias in favour of the more educated and people without children at home when compared to the general population. No statistically significant differences were found either at test or at retest between the eight experimental groups on the basis of their socio-demographic characteristics. Overall, there were similar numbers of respondents in the 8 groups although, as Table 3 shows, there were fewer in groups 6 and 8 at test and fewer in group 6 at retest.

At test 14 interviews were found to be incomplete in that either one or both of the main valuation methods was missed out entirely and these respondents have been excluded from subsequent analysis. Of these 14 respondents, 71% were aged 61 or over, none were in paid work (36% of them were retired) and 71% were women. Those with incomplete interviews also had more impairment than the rest of the respondents on all the Euroqol dimensions except for mood, where they reported levels of anxiety or depression similar to that of those with complete interviews. Reasons for incomplete interviews varied; some were elderly or confused, one was non-English speaking, one refused to "gamble with God", and another had difficulty in putting other states into perspective due to being in very poor health himself. All 110 re-interviews were complete.

3.2. Completeness

Table 4 shows that at both test and retest TTOP task was the most complete. In the test data, TTOP was significantly more complete than any of the other main methods (all at $p < 0.01$). In the retest data, TTOP was more complete than SGNP ($p < 0.05$), with *no* missing values.

3.3. Logical consistency

Given the ordinal structure of the component dimensions in the Euroqol descriptive system, some states are logically ordered with respect to others. For example, it would be expected that 21111 should be given a higher score (to

Table 2
Respondent characteristics

	Test ($n = 335$)		Retest ($n = 110$)		General population
	%	n	%	n	(GHS 1989)%
Female	58.5	196	54.5	60	52.0
Age					
16–20	2.4	8	2.7	3	7.7[a]
21–60	69.5	233	70.9	78	69.2
61+	27.8	93	26.4	29	23.1
(missing)	(0.3)	(1)	0	–	
Have children living with them	33.1	111	32.7	36	47.0
Main activity					
Paid work	43.9	147	42.7	47	59.5
Looking after home	25.1	84	30.0	33	–
Other	31.1	104	27.3	30	–
Education					
Left school at min. age	48.4	162	46.4	51	–
Training since school	29.6	99	29.1	32	–
Degree/prof. qualification	22.1	74	24.5	27	8.0
Cigarette smoker	34.6	116	30.9	34	30.0
Health status					
Problems with:					
Mobility	21.2	71	18.2	20	
Self care	3.6	12	1.8	2	
Usual activity	15.8	53	14.5	16	
Pain	34.1	114	27.3	30	
Mood	23.0	77	16.3	18	
Experience of illness					
Job looking after ill people	14.6	49	20.9	23	
Serious illness					
in self	27.2	91	31.8	35	
in family	36.7	123	48.2	53	
in others	32.8	110	34.5	38	
Experience of any dysfunctional states used in survey	64.5	216	70.9	78	

[a] From GHS 1990, figures are for ranges 16–19, 20–59 and 60+.

indicate less severity) than 21221 because it is better on at least one dimension and no worse on any of the other dimensions. For some pairwise comparisons, there are no a priori expectations of this kind, e.g. between 21111 and 11122. Where an a priori expectation holds, it is termed a logical consistency.

Table 3
Order in which methods presented to respondents (numbers in parentheses refer to incomplete interviews)

Experimental group	1st method	2nd method	Test *n*	Retest *n*
1	TTO NP	SG NP	44 (2)	16
2	SG NP	TTO NP	46 (3)	16
3	TTO NP	SG P	49 (1)	17
4	SG P	TTO NP	44 (2)	12
5	TTO P	SG NP	44 (1)	19
6	SG NP	TTO P	29 (0)	4
7	TTO P	SG P	46 (3)	14
8	SG P	TTO P	33 (2)	12
	Total		335 (14)	110
	SG – props		172	55
	SG – no props		163	55
	TTO – props		152	49
	TTO – no props		183	61

With the states used here, 12 such comparisons are possible. A calculation has been made of the number of logically consistent rankings made by each respondent, expressed as a percentage consistency rate. Because the number of possible pairwise comparisons drops substantially when a respondent fails to value a state, the data of those respondents with more than *one* missing value on the SG or TTO were considered to be unusable in the calculation of consistency rates. In addition, the few respondents who gave the same score to five or all six states on the same method were also excluded from this analysis. The distribution of consistency rates was highly skewed, with the majority of respondents having rates close to 100% and a few respondents having rates below 50%. For this reason, the median was chosen as the appropriate measure of central tendency. Table 5 shows that the TTO variants have higher consistency rates than SG variants both at test and at

Table 4
Completion of each method

Method	*n*	States unvalued			
		Test		Retest	
		%	*n*	%	*n*
SGP	55	5.3	52	2.4	8
SGNP	54	4.4	41	6.2	20
TTOP	49	0.8	7	0	0
TTONP	61	4.2	44	3.0	11

Table 5
Consistency rates by method: Median (and interquartile range)

Method	Test		Retest	
	n	Consistency	*n*	Consistency
SGP	136	83.8 (66.7–91.7)	45	83.3 (66.7–91.7)
SGNP	145	87.5 (62.5–95.8)	47	83.3 (58.3–100)
TTOP	145	91.7 (75.0–91.7)	48	91.7 (77.1–91.7)
TTONP	163	91.7 (66.7–100)	58	91.7 (66.7–100)

retest but there are no statistically significant differences between any of the four main methods.

Consistency rates on the VAS (not reported here) were the same across the eight experimental groups, suggesting that differences in consistency rates between main methods were not attributable to a response bias. Consistency rates for each of the main methods when they were done first i.e. immediately after the VAS, showed no statistically significant differences, either at test or retest. Also, there was little difference between test and retest consistency rates since subtracting each respondent's consistency rate at retest from their rate at test yielded a median difference of zero for all methods. With respect to respondent characteristics, it appeared that level of education and consistency rate were positively related, particularly for the SG variants where those with a minimum education had significantly lower consistency rates ($p < 0.05$ on both variants). With respect to possible interviewer effects, a few interviewers had respondents with lower than average consistency rates, but results were not affected when data from these interviewers were removed from the analysis. Similarly, no "learning effect" was identified when each interviewer's first three interviews they conducted were compared with their remaining interviews.

3.4. Valuation results

Since there were no differences found for any of the methods according to the order of presentation of the task or according to whether the preceding task was a

Table 6
Valuations for each state-test: Medians (and interquartile ranges)

State	SGP	SGNP	TTOP	TTONP
21111	0.85 (0.60–0.95)	0.90 (0.75–0.95)	0.95 (0.75–0.95)	0.95 (0.85–0.95)
11122	0.70 (0.45–0.90)	0.85 (0.50–0.90)	0.90 (0.70–0.95)	0.90 (0.65–0.95)
21221	0.60 (0.25–0.75)	0.75 (0.50–0.90)	0.80 (0.60–0.90)	0.85 (0.65–0.90)
21232	0.30 (0.15–0.55)	0.55 (0.30–0.80)	0.45 (0.05–0.75)	0.55 (0.30–0.70)
22323	0.35 (0.10–0.55)	0.50 (0.30–0.80)	0.40 (0–0.70)	0.55 (0.30–0.70)
33333	0.00 (−0.05–0.10)	0.10 (−0.10–0.40)	−0.30 (−2–0.05)	0.10 (−1.5–0.45)

Table 7
Valuations for each state: Retest medians (and interquartile ranges)

State	SGP	SGNP	TTOP	TTONP
21111	0.85 (0.55–0.95)	0.90 (0.70–0.95)	0.95 (0.75–0.95)	0.95 (0.80–0.95)
11122	0.70 (0.50–0.85)	0.80 (0.35–0.90)	0.90 (0.55–0.95)	0.80 (0.60–0.90)
21221	0.70 (0.40–0.85)	0.65 (0.45–0.85)	0.80 (0.60–0.90)	0.80 (0.60–0.90)
21232	0.35 (0.10–0.60)	0.50 (0.30–0.70)	0.40 (0.05–0.75)	0.60 (0.30–0.80)
22323	0.30 (0.05–0.50)	0.50 (0.25–0.80)	0.30 (0–0.65)	0.55 (0.30–0.70)
33333	0.00 (−0.05–0.10)	0.05 (−2–0.40)	−0.60 (−3–0.05)	0.05 (−4–0.40)

props or a no props variant, Table 6 shows the valuations for each health state (at test) from the four main methods. The predominant order of states is 21111, 11122, 21221, 21232, 22323, 33333 but SGP produces a "reversed" order for 21232 and 22323, although the valuations given to these two states are close together for all methods anyway. In general, it appears that the no props variants yield higher values than the props ones and that TTO values are higher than the SG ones although, interestingly, TTOP is the only method which gives a negative median score to state 33333. Table 7 shows the valuations elicited at retest where the predominant order of states is the same as that at test and again 33333 is, on average, considered to be worse than dead on TTOP.

Table 8 shows the results of a within-respondent comparison of valuations using the Wilcoxon matched-pairs signed-rank test. The results confirm those indicated in Tables 6 and 7, suggesting that: (1) TTONP values are significantly higher than SGP ones for all states except 33333, (2) SGNP values are higher than TTOP ones for the three most severe states, (3) TTOP values are higher than SGP ones for the three least severe states and lower for 33333, and (4) there are no significant differences between TTONP and SGNP valuations. There are fewer significant differences between methods at retest than at test due partly to the smaller number of respondents at retest.

Table 8
Within-respondent comparison of valuations

State	TTONP v SGNP	TTONP v SGP	TTOP v SGNP	TTOP v SGP
21111		T R		T R
11122		T		T R
21221		T R		T R
21232		T	T	
22323		T	T	
33333			T	X

T = TTO valuation is higher than SG one at test ($p < 0.05$).
R = TTO valuation is higher than SG one at retest ($p < 0.05$).
X = SG valuation is higher than TTO one at least ($p < 0.05$).

Table 9
Effect of own health state on valuations (figures are median scores). All differences shown are significant at $p < 0.05$ (Mann–Whitney U tests)

Method	State	Own health state dysfunctional [a]							
		Mobility		Usual activities		Pain/discomfort		Anxiety/depression	
		No	Yes	No	Yes	No	Yes	No	Yes
SGP	22323	0.35	0.45						
	33333				−0.05	0.05			
SGNP	21111							0.90	0.95
	11122			0.85	0.90			0.85	0.90
	21221	0.70	0.85	0.75	0.85				
	21232	0.50	0.70	0.50	0.70				
	22323	0.45	0.63	0.45	0.60			0.45	0.55
TTOP	21221	0.75	0.85	0.75	0.85	0.75	0.85		
	21232					0.40	0.55		
TTONP	33333	0.05	0.3						

[a] Self care omitted due to small numbers with any problems at all.

3.5. Construct validity

Construct validity in this paper relates to the background characteristics of respondents that are (and are not) expected to account for variance in valuations. The constructs tested here are that those in poor health should give higher valuations than those in good health but that valuations should not differ by any other background characteristic. Table 9 shows that respondents who were themselves in a dysfunctional health state (i.e. reported being in either level 2 or 3 on a dimension) did give significantly higher scores to some but not all states. Other background characteristics, such as age, gender and employment status, showed no systematic influence on valuations.

3.6. Test–retest reliability

The interval between the first and second interview varied from 6 to 16 weeks, (median of ten and a half weeks). At retest respondents were asked "Has anything important happened to you since the last interview a few months/weeks ago?". 29 of the 110 test–retest respondents (26%) reported that they had experienced an important event of whom all but three reported a deterioration in the own or someone else's health. As a group these people reported significantly more impairment of mobility and usual activities than the other respondents (both $p < 0.05$), and also reported more pain ($p < 0.01$) and anxiety/depression ($p < 0.05$). Reflecting this, they also reported more personal experience of illness ($p < 0.05$). Since this greater experience of very recent illness may affect the

Table 10
Mean correlation coefficients between test and retest scores

Method	Without important event n	With important event n	Spearman		Pearson	
			Without	With	Without	With
SGP	25	11	0.63 [a]	0.750	0.63 [a]	0.691
SGNP	31	8	0.71	0.529	0.74	0.498
TTOP	37	11	0.81	0.727	0.83	0.763
TTONP	25	14	0.54 [a]	0.643	0.55 [a]	0.622

[a] Significantly lower than TTO props ($p < 0.05$).

respondents' valuations of health states, those re-interviewed were separated into two groups on the basis of whether or not they reported that they had experienced an important event since the first interview.

Treating the data first as ordinal and then as cardinal, Spearman's rank coefficient and Pearson's r coefficient were calculated. The mean correlations for those without and with important events are shown in Table 10 which shows that TTOP has the highest correlation coefficients and for those without important events performs significantly better than both SGP and TTONP ($p < 0.05$). Table 11 shows the correlation coefficients for each method separated according to the time interval between test and retest. TTOP and SGNP have the highest correlations for respondents re-interviewed "early" i.e. within the median time interval of 73 days. While both the Spearman and Pearson correlations for SGNP fall as the time between test and retest increases, the corresponding values for TTOP remain at high levels.

In terms of median differences in scores between test and retest, there are no significant differences between test and retest for any state within any method for those *not* reporting important events. For those with an important life event, only two differences are significant at the 5% level (both on SGNP), suggesting that important life events have negligible effects on state-by-state valuations.

Table 11
Mean correlation coefficients for different time intervals between test and retest (for respondents without an important event)

Method	n		Spearman		Pearson	
	< 73 days	> 73 days	< 73 days	> 73 days	< 73 days	> 73 days
SGP	11	14	0.60	0.64	0.57	0.67
SGNP	19	12	0.79	0.56	0.83	0.59
TTOP	18	19	0.81	0.81	0.83	0.83
TTONP	13	12	0.54	0.48	0.64	0.42

4. Discussion

The main aim of this study was to select one method of valuing Euroqol health states by interview. Since a choice could not be made between SG and TTO on theoretical grounds or on the basis previous empirical work, the study was designed to allow a direct comparison between these two methods. Each method was tested in two variants; props and no props.

On the grounds of completeness, there is evidence in favour of TTOP since it was significantly the most complete of the main methods at test and had *no* missing data at retest. No clear "winner" emerged from a test of logical consistency but TTOP would be given a slight preference. This issue has rarely been considered in valuation studies and would be unimportant if all methods generated similar (high) levels of consistency. However, if consistency rates are low then doubt is cast on the feasibility of valuing health states in this way. Our experience here is that there is a "threshold" level of consistency of somewhere in the region of 85% for SG and 90% for TTO. We consider these rates to be acceptable.

The construct validity of the methods was assessed according to the extent to which valuations differed by the background characteristics that previous literature had shown to be important (and unimportant) determinants of valuations. It was hypothesised that valuations would not differ according to the age, gender and employment status of the respondent but that higher valuations would be elicited from respondents with experience of illness. All methods yielded valuations which supported the former construct whilst tentative support was lent to the latter construct. Of course, if the construct is not supported it does not necessarily invalidate the method as it may be that the construct itself is misspecified. More research is needed before the constructs hypothesised in this paper can be considered to be absolute standards.

An alternative way to assess validity is suggested by Nord (1991). He argues that the validity of any valuation technique will necessarily depend on how the values elicited are interpreted. Since, when combined with life expectancy to generate Quality-adjusted life-years (QALYs) they are often interpreted as measures of social value, Nord proposes testing validity by asking respondents whether they agree with the consequences in terms of the implied priorities for health care. He suggests that this should be done by comparing the valuations elicited by SG or TTO with those elicited from the equivalence of numbers procedure. However, this test of validity is predicated on the assumption that equivalence of numbers techniques represent the "gold standard" by which other methods are to be judged. But different methods measure different things. The valuations elicited by the SG and the TTO methods are based on assessments of *individual* utility whilst the valuations yielded by equivalence of numbers techniques are based on considerations of the utility gained by *other people* who receive treatment. This may be seen as an advantage in the context of health care

but the implications of it need to be well thought out. For example, an individual's health may be unimportant to her but very important to others. The issue of whose values should count in this context is unresolved. Therefore, it was decided not to test validity in this way.

Alternatively, the validity of health state valuations may be assessed by considering the extent to which the valuations elicited by these methods are valid representations of individual preferences. One way to test this would be to examine the robustness or otherwise of the valuations. Less confidence would be placed in valuations that are sensitive to seemingly irrelevant changes in problem structure or question format, for example. It is encouraging that valuations from *all* methods appeared to be unaffected by the order of presentation i.e. valuations were no different whether that task was administered first or second, or whether it was preceded by a props or a no props variant. In this way valuations from both SG and TTO can be seen to be insulated from (irrelevant) framing effects. This finding contradicts that of Llewellyn-Thomas et al. (1982) who found the existence of an anchoring effect when riskless methods such as the TTO were preceded by lottery questions.

It is also encouraging that all the methods produce a similar *ordinal* ranking of health states, which suggests that they all allow respondents to differentiate between states of differing severity. However, differences in *cardinal* values are observed. In general, valuations are highest when elicited using either of the no props variants, "intermediate" when using TTOP, and lowest when using SGP. These findings raise two important questions. First, why do no props variants yield higher valuations than the props variants? Second, why are TTOP valuations higher than SGP valuations?

The answer to the first question may lie in the different ways in which response categories were presented to respondents. In the props variants, respondents were presented with choices in a "ping-pong" fashion, moving back and forth between higher and lower probabilities of success in the SG and longer and shorter life expectancy in the TTO. In the no props variants, on the other hand, respondents were presented with all possible responses at once. These were listed from high to low probability of success in the SG and from long to short life expectancy in the TTO. It is likely that respondents would have started from the top of the page and worked their way down. It is possible that this may have resulted in a *reference point* effect in which respondents gave special status to favourable outcomes and hence to higher (inferred) health state valuations. This suggests that no props variants may introduce systematic bias into valuations i.e. may be susceptible to framing effects.

That TTOP yields higher valuations than SGP goes against much of the theoretical and empirical work which suggests the opposite relationship. It is postulated that if utility on the ordinate is plotted against length of life on the abscissa, the resulting utility function is concave to the origin. Two assumptions account for this concavity. First, it is assumed that people have positive time

preference in that they value years of life in the near future more highly than they value years of life in the more distant future (Gafni and Torrance, 1984). Second, it is assumed that people are risk averse and have an aversion to gambling, particularly to gambles involving life or death outcomes (Bombardier et al., 1982).

The time effect implies that people will be more willing to give up years of life at the end of a profile (as in the TTO) than they will be at the beginning of, or during, the profile. The gambling effect implies that people will be less willing to accept the gamble outcomes in the SG and more willing to accept the certain outcome. Therefore, the SG is expected to yield higher values than the TTO. This is borne out by the much of the evidence to date (Wolfson et al., 1982; Read et al., 1984). The results presented in this paper contradict this evidence and suggest that people are *more* willing to take an uncertain gamble involving the risk of dying than they are to sacrifice a certain amount of their life expectancy. These findings highlight the importance of more qualitative research into the cognitive processes involved in the formulation of individual preferences.

Whatever the reasons, it appears that valuations from TTOP are the most central in that they are generally higher than SGP ones and lower than SGNP and TTONP ones. The exception appears to be state 33333 which has a lower score on TTOP than on any of the other methods. Indeed, TTOP is the only method which results in a *negative* median value for this state. In other words, at least half the people valuing this state on this method consider it to be worse than death. This may be because the TTO method forces respondents to think more closely about the consequences of being in an extremely dysfunctional state for 10 years without any change. In the SG, the duration element may be given less prominence by respondents. There is evidence to suggest that more states are regarded as worse than death the longer they last (see Sutherland et al. (1982) who postulate the concept of a "maximal endurable time", after which some states yield a negative utility). In this respect, valuations to TTOP may more accurately represent individual preferences *for states that last 10 years without any change.* That some states may be considered as being worse than death is not unique; they have appeared in several countries for several valuation methods (Rosser and Kind, 1978; Read et al., 1984).

Before definite conclusions can be reached on the issue of which method most accurately represents individual preferences, it is important to gain a better understanding of the reasons *why* valuations differ (both within and between sub-groups). This issue is complicated by the fact that health state valuations from choice-based methods are likely to be a function of both the severity of the health state *and* the context of the choice. For example, responses to willingness-to-pay questions are likely to be influenced by initial levels of wealth and by the utility derived from money; responses to standard gamble questions are likely to be influenced by attitudes to risk; and responses to time trade-off questions are likely to be influenced by life expectancy and time preference.

Test–retest reliability gave a more definitive answer in that TTOP valuations

showed the most stability across time, performing significantly better than both TTONP and SGP and similarly to SGNP. It has been conventional in this field to assess test–retest reliability by calculating the correlation coefficient between the first and the second sets of scores obtained from each respondent. The coefficient of 0.81 for TTOP compares well with those from other studies (Churchill et al., 1984; O'Connor et al., 1985; Churchill et al., 1987), particularly as the time between test and retest was longer at a median of ten and a half weeks.

Bland and Altman (1986) have argued that use of correlation is misleading since it measures only the strength of a relation between two variables and not the agreement between them. Instead, they suggest plotting the differences in scores between two methods (or in this case the difference between test and retest scores) against their mean. By calculating "limits of agreement" between the two sets of scores (defined by Bland and Altman as the mean plus and minus two standard deviations) and the confidence intervals associated with them, the degree of agreement between the sets of scores can be summarised. However, given that six states were valued using four methods, there would be twenty-four graphical representations of the differences between test and retest scores. This would mean that unless one method produced the greatest agreement between test and retest for all six states (which is not the case), then it would be extremely difficult to determine the overall performance of each method. For this reason we feel that the correlation coefficient provides the best summary statistic available.

On the basis of the results reported here, TTOP has been chosen as the valuation method to be used in a large survey of the UK general population. It is recognised that no clear cut "winner" emerged from this study and, in particular, there is little to choose between TTOP and SGNP. The need to select one method, however, has pushed the balance in favour of TTOP as the method which performed significantly better on completeness, marginally better on logical consistency, and significantly better than SGNP and TTONP on test–retest reliability. In addition, the possibility that questionnaire framing may bias responses to the no props variants casts some doubt on the validity of the valuations from these methods.

4.1. Other issues

It should be noted that the choice of TTOP has been made in the context of a study conducted with a random sample from the British general population, using a particular descriptive tool for health status, and with specially designed boards and protocols. Although the method performed well, it was clear from interviewers' comments that improvements could be made to ease the handling of scripts, cards and boards in an often confined space. This would be particularly important in a clinical setting although it is encouraging that TTO has been found to be relatively easy in practice (Torrance, 1987) and has been used fairly widely to generate valuations for health states (Singer et al., 1991; Laupacis et al., 1992). It

is not known whether the choice of the Euroqol descriptive system affected the outcome, although it is unlikely to have had a differential effect on the SG and TTO methods.

Although the props versions of both SG and TTO were based on "classical" procedures (Furlong et al., 1989; Torrance, 1976; Torrance, 1986), both included modifications primarily to simplify the procedures for both interviewers and respondents (see Gudex (1994a) and Gudex (1994b) for details of the protocols). The changes to the McMaster Chance Board simply replaced a probability wheel with a sliding scale and therefore is unlikely to have had any systematic effect on valuations. The changes to the TTO protocol for states rated as worse than death may have had an effect on valuations since respondents may value a scenario in which a bad state is followed by a good state differently from one in which a good state is followed by a bad state, even though the time spent in each of the states may be identical. This is an empirical question which requires attention.

There is the general issue of how to score states worse than death. Given the standard health preference scale, states preferred to death are limited by an upper bound of one. However, there is no comparable lower bound for states worse than death which in this study could take a value as low as -19. Since our chosen measure of central tendency was the median, this asymmetry between positive and negative values had little effect on overall values even for the TTOP method which produced the most negative valuations. However, the implication of using the mean is that a minority of respondents rating a state as worse than death may more than offset a majority of respondents rating it as better than death. As Torrance (1984) noted "this issue of large negative values and what to do about them needs much more study".

Of significance to a future study is the failure here to obtain a representative sample of the general British population. Although the response rate was 55%, it is of concern that the respondents as a whole were more educated than the general population. This suggests the need for particular care over sampling procedures for a larger study.

5. For further reading

Brooks, 1991; Rawls, 1971; Thomas and Thomson, 1992.

Acknowledgements

The study was conducted in collaboration with Social and Community Planning Research and was financed by the Department of Health. We are grateful for the comments of our colleagues, Graham Loomes and Trevor Sheldon, and those of

two anonymous referees. We are also grateful for the secretarial and logistic support provided by Sally Baker and Kerry Atkinson.

References

Bennet, K., G.W. Torrance and P. Tugwell, 1991, Methodologic challenges in the development of utility measures of health related quality of life in rheumatoid arthritis?, Controlled Clinical Trials 12.

Berzon, R. and S. Shumaker, 1993, A critical review of cross national health-related quality of life instruments, Quality of Life Newsletter 5, 1–2.

Bland, J.M. and D.G. Altman, 1986, Statistical methods for assessing agreement between two methods of clinical measurement, Lancet i, 307–310.

Bombardier, C., A.D. Wolfson, A.J. Sinclair and A. McGreer, 1982, Comparison of three preference measurement methodologies in the evaluation of a functional status index, in: R.B. Deber and C.G. Thompson, eds., Choices on health care: Decision-making and evaluation of effectiveness.

Brazier, J., N. Jones and P. Kind, 1993, Testing the validity of the Euroqol and comparing it with the SF36 health survey questionnaire, Quality of Life Research 2, 169–180.

Brooks, R.G., 1991, Health status and quality of life measurement: Issues and developments, Swedish Institute for Health Economics, IHE, Lund.

Buckingham, K. and M. Drummond, 1993, A theoretical and empirical classification of health valuation techniques, HESG Conference, Strathclyde.

Bullinger, M., 1991, Quality of life: Definition, conceptualization and implications – a methodologist's view, Theoretical Surgery 6, 143–148.

Carter, W.B, R.A. Bobbitt, M. Bergner and B.S. Gibson, 1976, Validation of interval scaling; the sickness impact profile, Health Services Research, 516–528.

Churchill, D.N., J. Morgan and G.W. Torrance, 1984, Quality of life in end-stage renal disease, Peritoneal Dialysis Bulletin 4, 20–23.

Churchill, D.N., G.W. Torrance, D.W. Taylor, C.C. Barnes, D. Ludwin, A. Shimizu and K.M. Smith, 1987, Measurement of quality of life in end-stage renal disease: The time trade-off approach, Clinical and Investigative Medicine 10, 14–20.

Essink-Bot, M.L., 1990, Valuations of health states by the general public: Feasibility of a standardised measurement procedure, Social Science and Medicine 31, 1201–1206.

Euroqol Group, 1990, Euroqol: A new facility for the measurement of health related quality of life, Health Policy 16, 199–208.

Froberg, D.G. and R.L. Kane, 1989a, Methodology for measuring health state preferences – II: Scaling methods, Journal of Clinical Epidemiology 42(5).

Froberg, D.G. and R.L. Kane, 1989b, Methodology for measuring health state preferences – III: Population and context affects, Journal of Clinical Epidemiology 42(6).

Furlong, W., D. Feeny, G.W. Torrance, R. Barr and J. Horsman, 1989, Guide to design and development of health-state utility instrumentation, CHEPA Working Paper, McMaster University, Hamilton, Ontario, Canada.

Gafni, A. and G.W. Torrance, 1984, Risk attitude and time preference in health, Management Science 30.

Gafni, A., S. Birch and A. Mehrez, 1993, Economics, health and health economics: HYEs versus QALYs, Journal of Health Economics 11, 325–339.

Gudex, C., 1994a, Standard gamble user manual: Props and self-completion method, Centre for Health Economics, University of York, Occasional Paper Series.

Gudex, C., 1994b, Time trade-off user manual: Props and self-completion method, Centre for Health Economics, University of York, Occasional Paper Series.

Hornberger, J.C., D.A. Redelmeier and J. Petersen, 1992, Variability among methods to assess patients well-being and consequent effect on a cost-effectiveness analysis, Journal of Clinical Epidemiology 45(5), 505–512.

Kaplan, R.M., J.W. Bush and C.C. Berry, 1978, The reliability, stability and generalisability of a health status index, Social Statistics Section, American Statistical Association, 704–709.

Laupacis, A., N. Muirhead, P. Keown and C. Wong, 1992, A disease-specific questionnaire for assessing quality of life in patients on hemodialysis, Nephron 60, 302–306.

Llewellyn-Thomas, H., H.J. Sutherland, R. Tibshirani, A. Ciampi, J.E. Till and N.F. Boyd, 1982, The measurement of patients' values in medicine, Medical Decision-Making 2, 449–462.

Loomes, G.L. and R. Sugden, 1982, Regret theory: An alternative theory of rational choice under uncertainty, Economic Journal 92.

Martin, J. and D. Elliot, 1992, Creating an overall measure of severity of disability for the office of population censuses and survey disability survey, Journal of the Royal Statistical Society A 155(1), 121–140.

Mehrez, A. and A. Gafni, 1991, Quality-adjusted life years, utility theory, and health-years equivalents, Medical Decision Making 11(2).

Munro, S., B. Ferguson, E. Sutcliffe and A. Cooper, 1992, St James's University Hospital NHS trust: Health outcomes project, York Health Economics Consortium, University of York, England.

Nord, E., 1991, Methods for establishing quality weights for life years, Working Paper 8, National Centre for Health Program Evaluation.

O'Connor, A.M., N.F. Boyd and J.E. Till, 1985, Influence of elicitation technique, position order and test–retest error on preferences for alternative cancer drug therapy, Paper to 10th National Nursing Research Conference, University of Toronto.

Pliskin, J.S., D.S. Shepard and M.C. Weinstein, 1979, Utility functions for life years and health status, Operations Research.

Rawls, J., 1971, A theory of justice, Harvard University Press, Cambridge.

Read, J.L., R.J. Quinn, D.M. Berrick, H.V. Fineberg and M.L. Weinstein, 1984, Preferences for health outcomes: Comparison of assessment methods, Medical Decision Making 4(3), 315–329.

Rosser, R. and P. Kind, 1978, A scale of evaluations of states of illness: is there a social consensus?, International Journal of Epidemiology 7, 347–358.

Sackett, D.L. and G.W. Torrance, 1978, The utility of different health states as perceived by the general public, Journal of Chronic Diseases 31, 697–704.

Schoemaker, P.J.H., 1982, The expected utility model: Its variants, purposes, evidence and limitations, Journal of Economic Literature 20, 529–563.

Sculpher, M., S. Bryan, N. Dwyer, J. Hutton and G. Stirrat, 1993, An economic evaluation of transcervical endometrial resection versus abdominal hysterectomy for the treatment of menorrhagia, British Journal of Obstetrics and Gynaecology 100, 244–252.

Singer, P.A., E.S. Tasch, C. Stocking, S. Rubin, M. Sieglerer and R. Weichselbaum, 1991, Sex or survival: Trade-offs between quality and quantity of life, Journal of Clinical Oncology 9(2), 328–334.

Sutherland, H.J., H. Llewellyn-Thomas, N.F. Boyd and J.E. Till, 1982, Attitudes towards quality of survival: The concept of maximum endurable time, Medical Decision-Making 2, 299–309.

Thomas, R. and K. Thomson, 1992, Health related quality of life: Technical report, Joint Centre for Survey Methods, London.

Torrance, G.W., 1976, Social preferences for health states: An empirical evaluation of three measurement techniques, Socio-economic Planning Sciences 10, 129–136.

Torrance, G.W., 1984, Health states worse than death, Paper to Third International Conference on System Science in Health Care, Berlin.

Torrance, G.W., 1986, Measurement of health state utilities for economic appraisal: A review, Journal of Health Economics 5, 1–30.

Torrance, G.W., 1987, Utility approach to measuring health-related quality of life, Journal of Chronic Diseases 40, 593–600.

Torrance, G.W., Y. Zhang, D. Feeny, W. Furlong and R. Barr, 1992, Multi-attribute preference functions for a comprehensive health status classification system, CHEPA Working Paper 92-18, McMaster University, Hamilton, Ontario.

von Neumann, J. and O. Morgenstern, 1953, Theory of games and economic behaviour (Wiley, New York).

Wolfson, A.D., A.J. Sinclair, C. Bombardier and A. McGreer, 1982, Preference measurements for functional status in stroke patients: Inter-rater and inter-technique comparisons, in: R.L. Kane and R.A. Kane, eds., Values and long-term care (Lexicon Books).