Pergamon

Soc. Sci. Med. Vol. 44, No. 10, pp. 1519–1530, 1997

PII: S0277-9536(96)00271-7
0277-9536/97 $17.00 + 0.00

MAPPING VISUAL ANALOGUE SCALE HEALTH STATE VALUATIONS ONTO STANDARD GAMBLE AND TIME TRADE-OFF VALUES

PAUL DOLAN[1] and MATTHEW SUTTON[2]

[1]Department of Economics, University of Newcastle-upon-Tyne, Newcastle-upon-Tyne NE1 7RU, U.K. and [2]Centre for Health Economics, University of York, York YO1 5DD, U.K.

Abstract—Despite becoming increasingly common in evaluations of health care, different methods of quantitatively measuring health status appear to produce different valuations for identical descriptions of health. This paper reports on a study that elicited health state valuations from the general public using three different methods: the visual analogue scale (VAS), the standard gamble (SG) and the time trade-off (TTO). Two variants of the SG and TTO were tested: Props (using specially designed boards and cards); and No Props (using a self-completion booklet). This paper focuses on empirical relationships between health state valuations from the VAS and the (four) other methods. The relationships were estimated using Tobit regression of individual-level data. In contrast to *a priori* expectations, the mapping functions estimated suggest that differences are more pronounced across variant than across method. Furthermore, relationships with VAS scores are found to depend on the severity of the state: TTO Props valuations are higher than VAS responses for mild states and lower for more severe states; SG Props valuations are broadly similar to VAS scores over a wide range; and No Props responses are consistently higher than VAS valuations, particularly for more severe states. Explanations are proposed for these findings. © 1997 Elsevier Science Ltd

Key words—health status measurement, visual analogue, standard gamble, time trade-off

INTRODUCTION

Because resources are scarce, choices have to be made about how they are allocated. It is increasingly recognised that the impact on the health status of the population served should be an important consideration when making these choices. There are two main stages in the development of any measure of health status that is intended for these purposes. The first is to describe health status in terms of certain aspects or dimensions. There is a consensus emerging regarding which dimensions are relevant to measures designed to describe general health status [1]. The second stage is to value the health states described. As yet there is little agreement as to which of the various valuation methods is the preferred technique. The choice of method is important, however, as different methods (and even different variants of the same method) have been found to yield different sets of valuations for identical descriptions of health status.

This paper reports on an empirical study in which health state valuations were elicited from all respondents using three methods: the visual analogue scale (VAS), the standard gamble (SG) and the time trade-off (TTO) [2]. Two variants of the SG and TTO were tested: one using specially designed boards and cards (Props); and one using a self-completion booklet (No Props).

There has been considerable debate concerning the relative merits of the SG and TTO methods. Economists have tended to prefer the SG method, because of its foundations in von-Neumann-Morgenstern expected utility theory. However, the TTO shares a common theoretical foundation with the SG in utility theory generally, i.e. they both require people to sacrifice one thing they value (life expectancy and certainty, respectively) in order to gain another thing they value (quality of life in both cases), such that they are indifferent between the two states of the world. Since both methods make assumptions about individual preferences that have been shown to be too restrictive to allow them to act as perfect proxies for utility [3], it is difficult to make a choice between them on theoretical grounds alone. Against this background, a number of hypotheses have been put forward to explain why they often yield different results [4].

Because valuations from the VAS are elicited in a choiceless context, and thus do not require people to make trade-offs between different arguments in their utility function, the method is commonly regarded as theoretically inferior to the choice-based SG and TTO methods. However, the VAS has the practical advantages of being simpler to complete and cheaper to administer than either the SG or the TTO. Consequently, it is widely used in clinical and evaluative studies. This paper, then, considers whether there exists an empirical relation-

ship between the valuations from the VAS and those from different variants of the SG and TTO. If an algorithm can be found which maps VAS values onto SG and/or TTO ones and if, crucially, the relationship is stable, then it might be possible to elicit valuations via (cheap and simple) VAS methods and "convert" them into (theoretically superior) SG and/or TTO values. Furthermore, the nature of these algorithms may provide useful insights into why different methods yield different valuations.

Torrance [5] was the first to try to construct conversion curves to relate different valuation methods. In a comparison of mean VAS and TTO values for the same states elicited from the same people, he concluded that "the two techniques exhibit a systematic relationship [that] can be approximated by a number of different functions. Two that fit well... are a logarithmic function and a power function" (p. 134), although these relationships did not hold at individual level. However, Wolfson *et al.* [6] obtained somewhat different results. Having estimated linear relationships, they concluded that standard gamble values were much higher than VAS ones, with TTO values generally lying somewhere in between. Similar results were obtained by Read *et al.* [7]. In a more recent study, Hornberger *et al.* [8] found the TTO produced the highest mean values, the VAS the next highest, and the SG yielded the lowest mean values. This contradicts much of the earlier work and may be in part a function of the fact that, whilst other studies invoked hypothetical scenarios, this study elicited patients' valuations of their own health.

Whilst these studies have shown beyond doubt that different methods can be expected to yield different sets of valuations, it is questionable to what extent their results and conclusions are generalisable. This is because: (1) all were based on small samples of convenient populations (none consisted of more than 67 people); (2) the analysis was performed on aggregate- rather than individual-level data, thus making the choice between competing models more difficult as well as making inefficient use of the data; and (3) all the results were generated using ordinary least squares (OLS) regression analysis, which is inappropriate, given the (censored) nature of the data being analysed. In contrast, the results presented in this paper are from a large-scale study of the general population, and are based on individual-level analysis using a Tobit model that takes account of the type of data typically dealt with in the health state measurement field.

HYPOTHESES

Despite their limitations, the results of earlier studies, together with some theoretical work, suggest two hypotheses. First, for identical descriptions of health status, valuations from the VAS will be lower than those from the SG or TTO. This may result from respondents' use of different *reference points* in the valuation of health states on the VAS than on the SG and TTO. For example, in the VAS, a respondent may take full health as their reference and value dysfunctional health states as losses from this state. In a qualitative study, Morris and Durand [9] suggest that VAS responses are indeed constructed in this way. In the SG and TTO, on the other hand, the respondent is asked to imagine already being in a dysfunctional health state and, consequently, this state becomes the reference point.

It seems entirely plausible that the value given to a particular health state will be a function both of its severity *and* of the state that it is viewed from, just as our perception of how fast a car is travelling is a function both of its actual speed *and* the speed it had previously been travelling at. This idea is developed by Kahneman and Tversky [10], who propose that an individual's value function is concave with respect to gains, convex with respect to losses, and steeper at each level of loss than at the corresponding level of gain, as depicted in Fig. 1. A (non-stochastic) value function with these properties will result in lower valuations from the VAS than from either the SG or the TTO.

Second, based on two assumptions about individual preferences, mapping functions for SG variants are expected to be different from those for TTO variants. The first assumption is that people are risk averse: if utility on the ordinate is plotted against length of life on the abscissa, the resulting utility function is concave to the origin. This implies that people will be less willing to accept the gamble outcomes in the SG and more willing to accept the certain outcome. The second assumption is that people have positive time preference: they value years of life in the near future more highly than they value years of life in the more distant future. This implies that people will be more willing to give up years of life at the end of a profile, as in the TTO. Thus, both assumptions imply that, for the same health

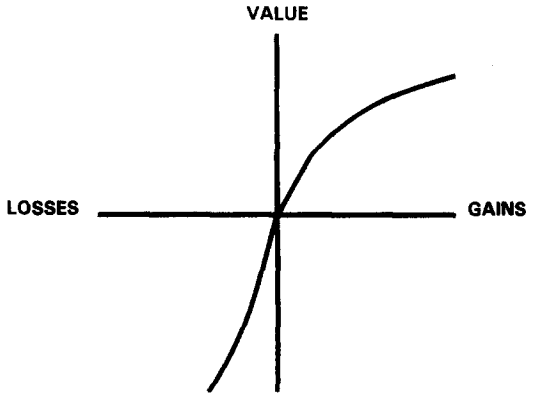

Fig. 1. A hypothetical value function.

states, SG values will be higher than TTO ones. Significant differences across variants of the same method are not anticipated.

STUDY DESIGN

The sample

The sample was drawn from adults aged 18 and over in the general population. A random sample of 700 addresses was drawn from 13 regional areas in the U.K. by Social and Community Planning Research (SCPR) using the postcode address file. The main fieldwork was carried out by 25 specially trained interviewers between March and May 1992. In order to study the two variants of the SG and TTO (from now on referred to as P and NP, to denote Props and No Props, respectively) and to test whether the order of presentation of these tasks influences valuations, each interviewer was randomly allocated to one of eight experimental groups (2 methods × 2 variants × 2 orders of presentation).

Of the 700 addresses selected for sampling, 88 (13%) were found to be "out of scope", being non-residential, empty/derelict, untraceable, or not yet built. Of the remaining 612 addresses, 335 interviews were achieved, giving a 55% response rate on in-scope addresses. The main reason for unsuccessful interviews was a refusal by the selected person. Table 1 shows that the sample was similar to the general population in terms of age and sex, although there was some response bias in favour of the more educated. Table 2 shows that, by chance, more respondents were in the four groups containing the TTONP variant than in the four groups containing TTOP. No statistically significant differences were found between the groups on the basis of their sociodemographic characteristics.

Table 2 also shows that 14 interviews were incomplete (defined as one or both of the main valuation methods being missed out entirely), and these respondents have been excluded from subsequent analysis. These respondents were scattered amongst the eight experimental groups and did not differ in terms of background characteristics from the remainder of the sample. Finally, Table 2 shows that the number of missing observations from the remaining respondents is very small, particularly for TTOP.

Table 1. Respondent characteristics

Characteristic	Sample (n = 335)		GHS[a] (1989)
	n	%	%
Female	196	58.5	52.0
Age: 16–20	8	2.4	7.7
21–60	234	69.8	69.2
61+	93	27.8	23.1
Educated to degree level	74	22.1	8.0

[a]The General Household Survey (GHS) is a nationally representative survey of households in Britain which is undertaken each year by the Office of Population Census and Surveys (OPCS).

Table 2. Completion of each method

Method	n	Exclusions	Missing values
SGP	172	8	5.3
SGNP	163	6	4.4
TTOP	152	6	0.8
TTONP	183	8	4.2

The questionnaires

All interviews used the EuroQol Descriptive System [11], which describes health status in five dimensions with no disease specificity. Previous piloting indicated that when asked to value the same health states using three methods, respondents could not effectively assess more than six states on each method (plus two anchor states on the VAS: full health and death). The states were chosen to be widely spread over the valuation space and, in the SG and TTO tasks, were presented to respondents in a standard order.

Each respondent was first asked to describe their own health using the EuroQol descriptive system. They were then asked to rank all eight health states. It was explained that each state was to be regarded as lasting for 10 years without change, followed by death. The respondent was asked to indicate where they would rate their own health on a vertical VAS, with end-points of 100 (best imaginable health state) and 0 (worst imaginable health state). They were then asked to rate the eight health states on an identical VAS. Once full health and death had been removed, the remaining six states (which were always presented in the same order) were valued using one variant of the SG and one variant of the TTO. At the end of the interview, personal background data were collected from each respondent.

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. For a state, h_b, rated as better than dead, the intermediate state is h_b and the treatment outcomes are full health and death, respectively. For a state, h_w, rated as worse than death, the intermediate state is death and the treatment outcomes are full health and h_w. In both cases, the object is to find the probability, p, at which the respondent is indifferent between the two alternatives.

For SGP, a sliding scale on a specially designed board showed the varying chances of success and failure of treatment. For each health state, the respondent was initially asked to choose between living for 10 years in that state and a treatment which would return them to full health for 10 years (i.e. a 100% chance of success). Then, to determine whether the state was considered to be better or worse than death, the respondent was asked to choose between 10 years in that state and a treatment which would result in immediate death (i.e. a 0% chance of success). If they preferred the former,

the protocol for states rated as better than dead was used and if they preferred the latter, the protocol for states rated as worse than dead was used. In both cases, the chances of success were presented in intervals of 10% in a "ping-pong" fashion, i.e. 90% success, 10% success, 80% success, etc. The question was complete when either (1) preferences changed over a 10% interval (e.g. the treatment was preferred when it had a 70% chance of success but the certain health state was preferred when the treatment had a 60% of success), in which case the state would be valued at halfway between the two probabilities (i.e. 0.65 in this example), or (2) indifference was reached.

The SGNP variant consisted of a self-completion booklet which showed the (certain and uncertain) alternatives on the lefthand page and the chances of success and failure relating to the uncertain treatment on the righthand page, which were listed in 10% intervals ranging from a 100% chance of success at the top to a 0% chance of success at the bottom. The respondent was initially shown the protocol for states rated as better than dead and was asked to place: a tick alongside all those probabilities where they would prefer the treatment; a cross alongside all those probabilities where they would prefer the certain health state; and an equals sign alongside the probability at which they would find it hardest to choose between the certain state and the treatment. If the treatment was preferred when it was certain to result in immediate death (i.e. when a tick was placed alongside a 0% chance of success), the respondent turned over the page and was presented with the protocol for states rated as worse than dead, which was competed in a similar fashion. Details of both SG protocols can be found in [12].

The TTO asks the respondent to choose between two alternatives. For a state, h_b, rated as better than dead, the first alternative is to live for a defined period of time, t, in h_b and then die. The second alternative is to live for a shorter period of time in full health and then die. For a state, h_w, rated as worse than dead, the first alternative is to die immediately and the second alternative is a number of years in h_w followed by a number of years in full health (which combined sum to t). In both cases, the time in full health, x, is varied until the respondent is indifferent between the two alternatives.

For TTOP a sliding scale on a double-sided board showed the number of years spent in each alternative: one side was used for states rated as better than dead and the other for states rated as worse than dead. Using the former side of the board, for each health state, the respondent was first asked to choose between living for 10 years in that state and living for 10 years in full health. Then, to determine whether the state was considered to be better or worse than death, the respondent was asked to choose between 10 years in that state and immediate death. If they preferred the former, they continued to use the side of the board for states rated as better than dead and if they preferred the latter, the side of the board for states rated as worse than dead was used. In both cases, the number of years spent in full health were presented in units of one year in a "ping-pong" fashion, i.e. nine years, one year, eight years, etc. The question was complete when either (1) preferences changed over a one-year period (e.g. full health was preferred when it was for seven years but 10 years in the particular state was preferred when full health lasted for six years), in which case the state would be valued at halfway between the two lengths of time (i.e. 0.65 in this example), or (2) indifference was reached.

The TTONP variant consisted of a self-completion booklet which showed the two health profiles on the lefthand page and the number of years spent in the profiles on the righthand page, which were listed in units of one year, ranging from 10 years in full health at the top to zero years in full health at the bottom. The respondent was initially shown the protocol for states rated as better than dead and was asked to place: a tick alongside the cases where they preferred a certain number of years in full health to 10 years in the particular state; a cross alongside the cases where they preferred 10 years in the particular state to the number of years in full health; and an equals sign alongside the case where they considered a certain number of years in full health to be equivalent to 10 years in the particular state. If immediate death was preferred (i.e. when a tick was placed alongside zero years in full health), the respondent turned over the page and was presented with the protocol for states rated as worse than dead, which was competed in a similar fashion. Details of both TTO protocols can be found in [13].

METHODS

Recalibration of scores

For both interpersonal and inter-method comparisons, it is necessary to recalibrate the raw scores elicited from each of the valuation methods. On all valuation methods, full health has been set equal to 1.0 and death has been given a value of 0.5. Although these two states are assigned the same value on all methods, it is possible for any other state to have a value that implies it is considered as bad as, or worse than, dead on one method yet to be rated as better than dead on another method. The value for death departs from the normal value of zero used in the literature so that mathematical manipulations of the VAS scores are well-behaved functions, and so that the SG and TTO scores can

be transformed onto a logarithmic scale for the analysis of power functions. Diagnostic tests for heteroskedasticity and functional form are included in the analysis, which should indicate any problems with this approach.

In cost–utility analysis, aggregation across respondents is achieved by measuring all individual valuations on a common 0–1 (dead to healthy) scale. Because respondents could locate full health and death anywhere on the VAS, it is necessary to recalibrate raw VAS scores so that "the unit of health" is the same across all respondents. Denoting the raw VAS scores for full health, death and some dysfunctional health state h by v_f, v_d and v_h, respectively, a health state index, h, can be derived according to the following decision rule:

if $v_f \leqslant v_h \quad h = 1$

if $v_d < v_h < v_f \quad h = 0.5 + [0.5^*(v_h - v_d)/(v_f - v_d)]$

if $v_d - (v_f - v_d)] < v_h < v_d \quad h = 0.5 - [0.5^*(v_h - v_d)/(v_f - v_d)]$

if $v_h \leqslant [v_d - (v_f - v_d)] \quad h = 0$

Both the SG and TTO implicitly assign full health and dead values of one and zero, respectively. Using notation introduced above, $h_b = p$ on the SG and x/t on the TTO, whilst $h_w = -p/(1-p)$ on the SG and $-x/(10-x)$ on the TTO. Negative scores do not have the same linear properties as those displayed by positive scores. In this study, valuations for states worse than dead have been transformed such that $h_w = -p$ on the SG and $(x/10) - 1$ on the TTO. This is a transformation that has been used elsewhere in the literature [14]. All SG and TTO scores were then recalibrated so that death equals 0.5 for each respondent.

Independent variables

Previous studies of the relationship between scores elicited by different techniques have concentrated on the relationships between aggregate (mean) scores. However, an alternative approach, for which a larger amount of data are available, is the estimation of mapping functions based on the individual-level data. Initially, a model using all the data was estimated which contained binary variables for variant (P or NP) and method (SG or TTO). This model failed the specification tests outlined in the next section, and thus the relationship is estimated separately for both variants. This produces four mapping functions to be estimated. The aim of these functions is to show the relationship between the VAS and the variants of the SG and TTO, irrespective of the health state being valued.

With SG and TTO scores as regressands, the regressors of interest are different transformations of VAS scores. However, in the analysis of the individual-level data, information is available for a number of respondent background characteristics which have been found to significantly affect health state valuations (for a review of the literature to date see [15]). If these independent variables affect scores derived from the methods differently, then these factors should be taken into account in estimating any mapping function. The characteristics considered in this analysis are sex, age (considered by the introduction of two dummy variables representing three age groups: 18–29 years; 30–59 years; and 60 years or more), and the individual's rating of their own health status on the VAS. Each of these variables is entered as an independent factor and as a cross-product term with the visual analogue score. Details of the abbreviations used for the independent variables are given in Table 3.

Models

Five functional forms were estimated to represent the relationship between VAS and SG/TTO scores: a linear, a quadratic and a cubic model plus two log-linear models. The linear model included only the VAS score, (V_{ix}), and the other independent variables. In the quadratic model $(V_{ix})^2$ was added, and in the cubic model $(V_{ix})^3$ was included:

Model 1: $Y_{ix} = \alpha + \beta V_{ix} + \theta_k \mathbf{Z}_{ik} + \varepsilon_{ix}$

Model 2: $Y_{ix} = \alpha + \beta V_{ix} + \delta V_{ix}^2 + \theta_k \mathbf{Z}_{ik} + \varepsilon_{ix}$

Model 3: $Y_{ix} = \alpha + \beta V_{ix} + \delta V_{ix}^2 + \rho V_{ix}^3 + \theta_k \mathbf{Z}_{ik} + \varepsilon_{ix}$

in which Y_{ix} is the recalibrated score on the SG/TTO for state x from individual i, V_{ix} is the score for health state x elicited from individual i by the visual analogue method, $\mathbf{Z}_{ik}$ is a vector of k independent variables for individual i, and ε_{ix} is an error term.

Two log-linear models were estimated:

Model 4: $Y_{ix} = A V_{ix}^{\beta} \mathbf{Z}_{ik}^{\theta k} \varepsilon_{ix}$

Model 5: $(1 - Y)_{ix} = A(1 - V_{ix})^{\beta} \mathbf{Z}_{ik}^{\theta k} \varepsilon_{ix}$

The difference between Models 4 and 5 relate to the testing of two assumptions: (1) a health state rated as good as full health on the VAS will also be rated as good as full health on the SG/TTO; and (2) a health state rated as far below dead as full health is above dead on the VAS will also be rated as far below dead as full health is above dead on the SG/TTO. Model 4 makes the first assumption, whilst the size of the constant in the regression can be used to test for the second assumption. Model 5 makes the second assumption, whilst the first can be tested from the regression results. In the estimation of his power function, Torrance [5] assumed equality between valuations from different methods at both end-points of the scale.

Estimation and testing

To take account of the fact that valuations on both the SG and the TTO were within the range 0.025–0.975, a Tobit estimation was undertaken

Table 3. Definition of variables

Variable	Definition
YOUNG	A dummy variable taking a value of 1 if the respondent is aged 18–29, and 0 otherwise
OLD	A dummy variable taking a value of 1 if the respondent is aged 60 + , and 0 otherwise
SEX	A dummy taking the value of 1 if the respondent is female, and 0 otherwise
SELF	The respondent's rating of their own health on the VAS
YOUNG*VAS	The product of the variable YOUNG and the VAS score for state x given by this respondent
OLD*VAS	The product of the variable OLD and the VAS score for state x given by this respondent
SEX*VAS	The product of the variable SEX and the VAS score for state x given by this respondent
SELF*VAS	The product of the variable SELF and the VAS score for state x given by this respondent

with censoring at both the top and bottom ends. The models were estimated using maximum likelihood within LIMDEP [16].

Two specification tests have been included in this analysis: a modified RESET test, and a test for heteroskedasticity in the error terms. The modified RESET test is undertaken in a two-step process. In the first stage, the model is estimated and the linear function is calculated:

$$\Gamma_{ix} = \beta_j \mathbf{X}_{ixj}$$

in which: βj are the j coefficients estimated in the Tobit model, where j refers to all independent variables in the model; and $\mathbf{X}_{ixj}$ is the vector of all independent variables which includes value(s) on the VAS. In the second stage, the square of the linear function is added to the equation, and the t-statistic on this variable can be used as a test of the functional form of the original model. The test for heteroskedasticity is created by a similar two-stage process. From the first stage, a new variable is created by taking the value of the linear function away from the actual value. A general test for heteroskedasticity is created, by regressing the square of this new variable on a constant term and the square of the linear function. The significance of the squared linear function term can be used as a test for general heteroskedasticity, which may also indicate general misspecification.

The specification of each regression equation was tested using these diagnostic checks. Any indicated problems with heteroskedasticity in the models were approached by estimating a Tobit model adjusted for heteroskedasticity. The variance of the error terms was assumed to be dependent on either the VAS (V_{ix}), the square of the VAS $(V_{ix})^2$, or the square of the linear function from the unadjusted Tobit estimation $(\Gamma_{ix})^2$.

Calculation of mapping functions

Because the data was analysed using a Tobit model, the predicted values from the regressions have to be transformed according to the following equations [16]:

$$E[y_i] = L\Phi_L + U(1 - \Phi_u) + (\Phi_u - \Phi_L)\beta_j \mathbf{X}_{ixj} + \sigma_i(\phi_L - \phi_u)$$

in which: L and U are the lower and upper bounds of the range within which y_i are restricted; β_j is the corresponding vector of coefficients for the set of independent variables $\mathbf{X}_{ixj}$; μ_i are the standard errors of the (possibly heteroskedastic) error terms; Φ is the standard normal cumulative distribution; ϕ is the standard normal distribution; and

$$\Phi_j = \Phi[(j - \beta_j \mathbf{X}_{ixj})/\sigma_i]$$

and

$$j = \phi[(j - \beta_j \mathbf{X}_{ixj})/\sigma_i]$$

in which: $j = L, U$. In contrast to OLS methods, the predicted values calculated from these equations are constrained within the desired range $[L,U]$.

RESULTS

The results of the various linear, quadratic and cubic models for each of the four variants are shown in Table 4. Where possible, results are shown for models which do not show evidence of functional form misspecifications or heteroskedastic disturbances at the 99% level. If all estimations fail to pass either or both of these tests, results are given for the Tobit estimations which are not adjusted for heteroskedasticity, unless one of the adjustments offers significant improvement on the original estimation. Results of Tobit models with (successful) adjustment for heteroskedasticity are shown for all valuation methods except SGNP, for which no adjustment made any significant improvement in this regard.

Few of the independent variables associated with respondent background characteristics are significantly different from zero at even the 5% level. For example, there is little evidence of any systematic effect of gender or self-rated health on the difference between scores from the VAS and the other methods. However, for both SG variants, the negative coefficient on YOUNG and the positive coefficient on YOUNG*VAS are both significant at the 5% level and, although not statistically significant, the sign on these coefficients is the same for the TTO variants. These results imply that, given the same VAS score, younger respondents tend to give lower valuations to more severe health states on the choice-based methods, particularly the SG. This divergence between age groups, however, decreases as the VAS rating increases.

Table 4. Regression results (*t*-statistics in parentheses)

Variable	TTO Props	TTO No Props	SG Props	SG No Props
V_{ix}	−2.86 (−3.81)	0.91 (3.59)	−1.51 (−2.29)	1.17 (3.51)
$(V_{ix})^2$	7.72 (6.02)		4.35 (4.25)	0.28 (1.74)
$(V_{ix})^3$	−4.30 (−5.58)		−2.50 (−4.46)	
YOUNG	−0.18 (−1.52)	−0.05 (−0.64)	−0.10 (−2.14)	−0.24 (−2.87)
YOUNG*VAS	0.17 (1.01)	0.14 (1.15)	0.27 (2.76)	0.32 (2.80)
OLD	−0.10 (−0.85)	0.39 (6.10)	0.13 (1.74)	0.10 (1.08)
OLD*VAS	0.01 (0.04)	−0.51 (−5.95)	−0.18 (−1.85)	−0.17 (−1.48)
SEX	−0.12 (−1.27)	0.04 (0.68)	0.07 (1.02)	−0.07 (−0.94)
SEX*VAS	0.18 (1.40)	0.02 (0.29)	−0.03 (−0.31)	0.14 (1.56)
SELF	0.15 (0.50)	0.10 (0.49)	−0.06 (−0.32)	0.83 (3.82)
SELF*VAS	−0.19 (−047)	−0.04 (−0.01)	0.05 (0.18)	−1.16 (−4.00)
Constant	0.46 (1.56)	0.01 (0.03)	0.49 (2.20)	−0.23 (1.11)
HET				
V_{ix}	0.18 (0.97)			
$(V_{ix})^2$		0.06 (0.83)	0.18 (1.55)	
$(V_{ix})^3$				
Functional form	(0.82)	(−1.38)	(−0.17)	(−1.54)
Hetero. disturbances	(−2.55)	(−2.11)	(−0.06)	(−3.22)
Likelihood ratio	0.18	0.18	0.28	0.37
Sample	855	982	836	869

The results for the two log-linear models are not very encouraging, and hence the results are not presented here. For the logarithmic model (Model 4), the likelihood ratio indices are below 0.05 for all methods, there is evidence of heteroskedastic disturbances in all four equations, and the estimated models for all but TTONP fail functional form tests. Similarly, with regard to the estimation of a power function (Model 5), the likelihood ratio indices are less than 0.01 for all methods, there is evidence of heteroskedasticity in all simple and adjusted equations, and no models pass functional form tests at the 5% level. The estimated likelihood ratio indices illustrate that the ability of power functions to explain variations in this data set is negligible. Moreover, although unlikely to be valid, the estimated coefficients on the constant terms are not close to one, as assumed by Torrance [5].

Mapping functions can be derived based on the predicted value formulae described above, a simulated index of possible VAS scores, and the average values of other independent variables. The average values of the other independent variables for which the predicted value functions have been estimated are given in Table 5. Average values of even the discrete variables are used, since this may reflect the characteristics of a general sample of respondents (e.g. the TTO Props case in which 29% are aged under 30 years and 20% are aged over 60 years). In contrast to the log-linear models, the linear functional forms pass misspecification and heteroskedasticity tests at the 99% level of significance for the majority of methods and, as indicated by the likelihood ratio index, explain a significant proportion of the variation in the data. As a result, mapping functions have been generated based on the coefficient estimates from the linear models only. These functions are shown in Figs 2–5.

As well as highlighting systematic differences across methods and variants, the mapping functions suggest that the relationship with VAS scores is a function of the severity of the state. Figures 3 and 5 both show that values from the No Props variants are higher than those for the VAS and that this difference increases as the severity of the state increases. Figure 2 shows that TTOP valuations are higher than VAS ones for mild states and lower for more severe states, although SGP valuations are broadly similar to VAS ones for a wide range of scores (see Fig. 4). Note that the functions estimated for the Props variants imply that for states worse than dead, a marginal increase in the VAS score is associated with a marginal decrease in the corresponding TTO or SG score. This finding is counter-intuitive, and thought to be related to both the lack of observations in this quadrant and the dominating effect of the function in other parts of the valuation space.

DISCUSSION

This paper has used health state valuation data from a large-scale general population study to estimate an empirical relationship between VAS scores and scores elicited from two variants of two choice-based methods. The analysis was based on individual-level data using the Tobit model, which takes account of the (censored) nature of the data. A number of different functional forms were tested

Table 5. Values of the independent variables used in the mapping functions

Variable	TTO Props	TTO No Props	SG Props	SG No Props
YOUNG	0.29	0.19	0.24	0.20
OLD	0.20	0.24	0.25	0.25
SEX	0.59	0.55	0.56	0.60
SELF	81.05	83.64	84.70	81.06

Note: The values of the variables YOUNG*VAS, OLD*VAS, SEX*VAS and SELF*VAS depend on the VAS score. The values for these variables at which the predicted value functions are evaluated are given by the product of the values given above and the particular value of the index function for the VAS.

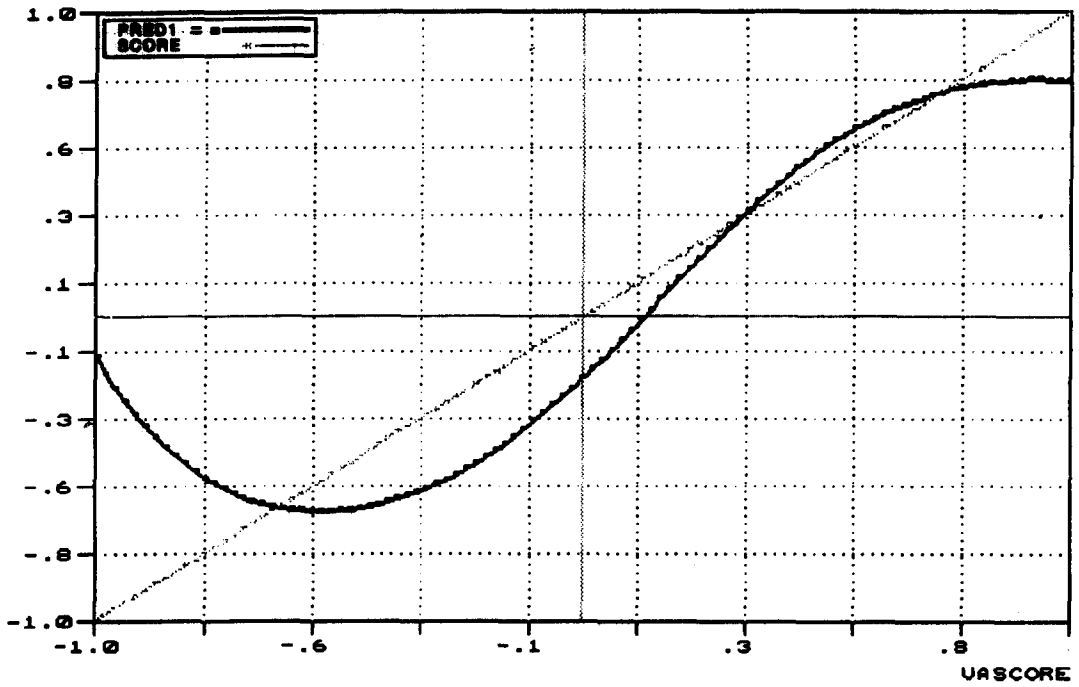

Fig. 2. Mapping function for TTO Props.

and a range of diagnostic tests was applied to the competing models. Logarithmic and power function formulations were outperformed by more flexible (linear, quadratic or cubic) functional forms, both in terms of specification and ability to explain variations in the data. In particular, the results do not lend support to the hypothesis that VAS and SG/TTO values can be related by a concave power function, since this functional form failed all diagnostic tests.

In contrast to expectations, differences between the mapping functions are more pronounced across variant than across method. That valuations differ by method variant is an important finding which

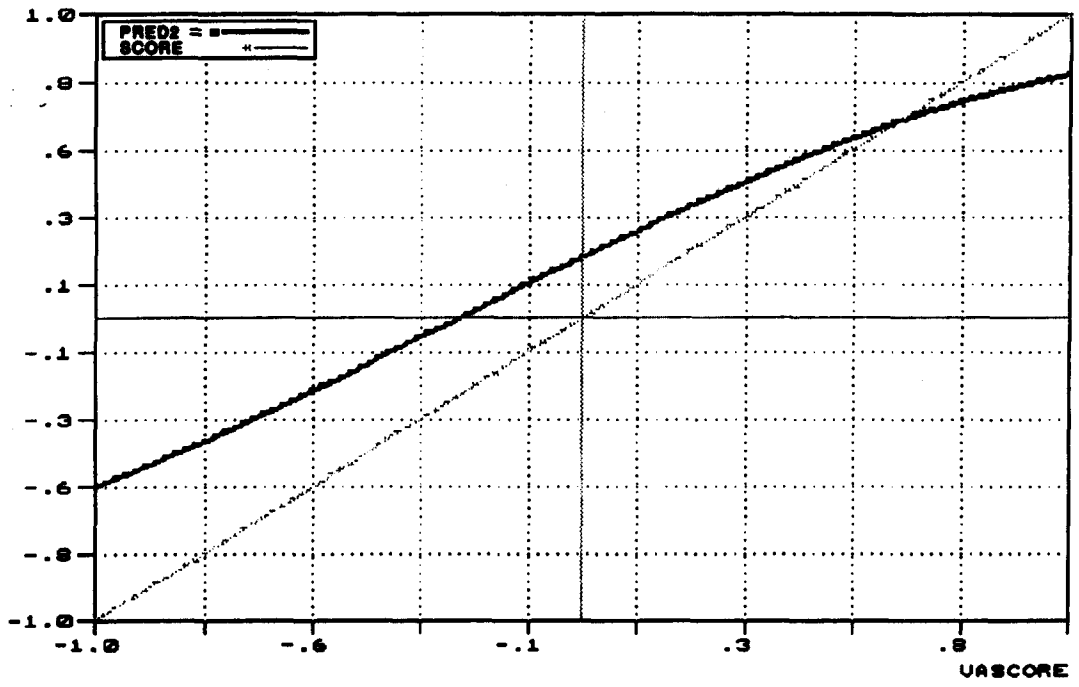

Fig. 3. Mapping function for TTO No Props.

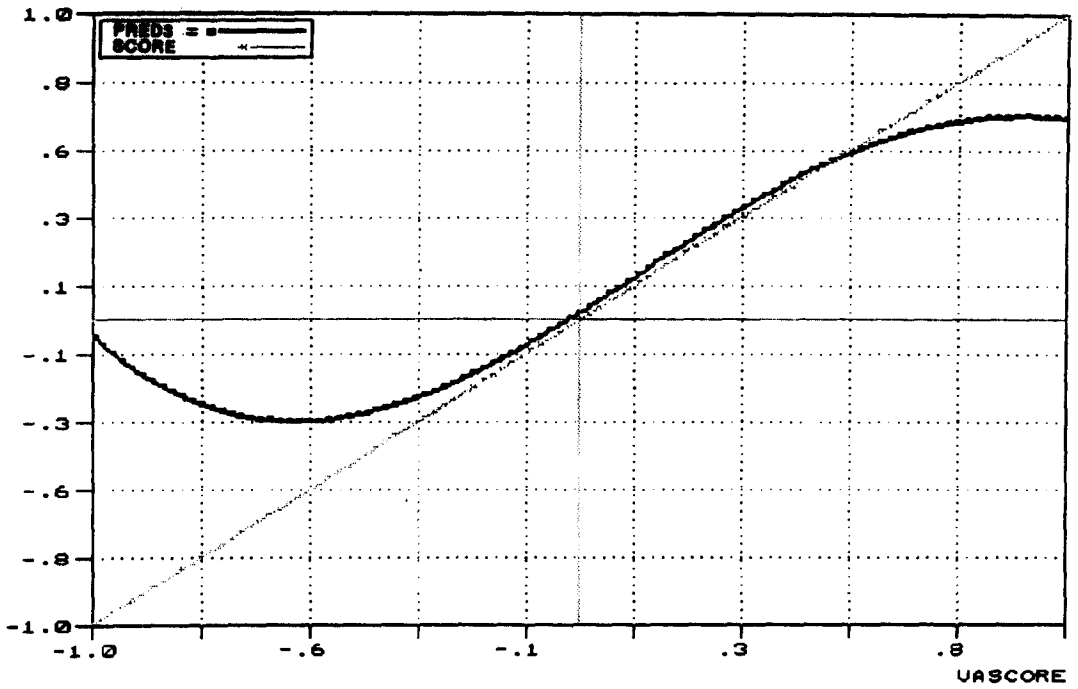

Fig. 4. Mapping function for SG Props.

has been noted elsewhere [17]. This result offers an explanation for the lack of consensus regarding the comparability of the different methods, since although the studies referred to in this paper used the same methods, the way in which they were administered differed enormously. This suggests that comparisons can only be made between valuations from different studies if both the same method *and* the same variant were used.

One of the underlying hypotheses was that, for identical descriptions of health status, VAS valuations would be lower than SG/TTO ones. Thus, for high valuations, the intercept of any mapping function between VAS and SG/TTO scores is expected

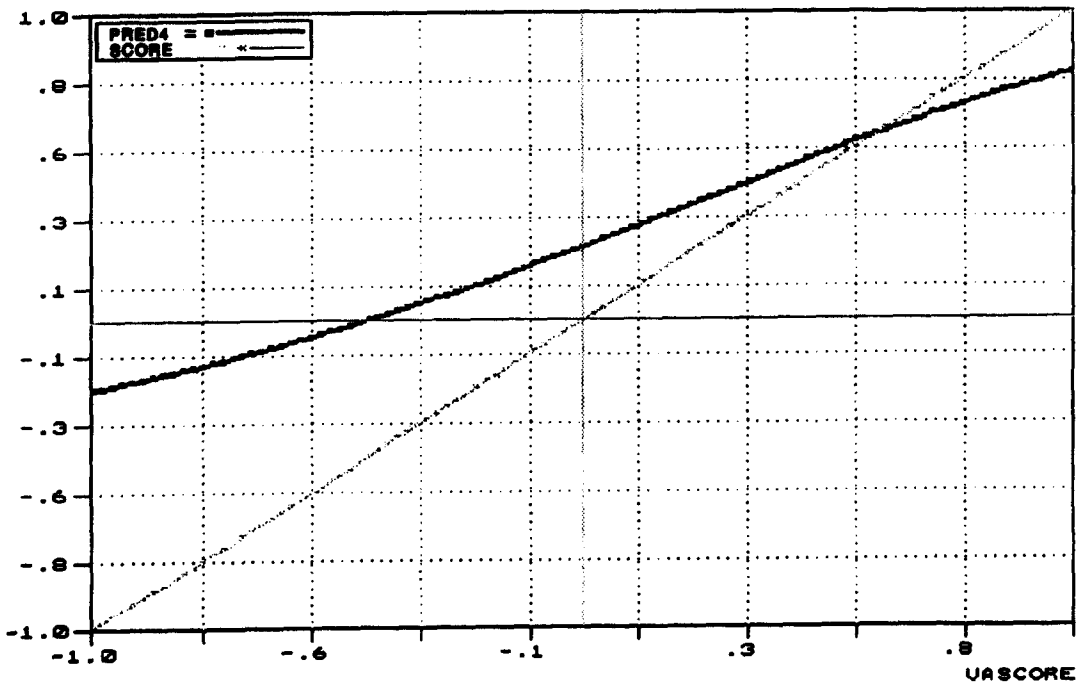

Fig. 5. Mapping function for SG No Props.

to be greater than one, whilst the gradient will be less than one. The gradients of all mapping functions for the linear models are indeed shallower than the 45° line. However, the mapping functions imply that a VAS score greater than about 0.8 is associated with a *lower* SG and TTO score. This suggests that the disutility associated with the very mild health states is *greater* for methods in which the reference point is the dysfunctional health state itself than for a method in which the reference point is likely to be full health. This result appears to cast doubt on the value function hypothesised by Kahneman and Tversky's Prospect Theory [10].

However, inspection of the distribution of health state values suggests that the majority of respondents *do* behave in accordance with Prospect Theory, i.e. have SG and TTO values that are higher than VAS ones for very mild states. Therefore, the fact that mapping functions cross the 45° line at about 0.8 appears to be explained by a small number of "outliers", who give high VAS scores yet give very low SG/TTO scores. Although there were no grounds for excluding such responses from the data set, this finding suggests that the functions presented here should be treated with caution, and their interpretation made clear. Because of the estimation technique used, the results represent mapping functions that show how *mean* VAS scores can be converted into *mean* SG and TTO ones.

Another unexpected result (at least in the presence of a reference point effect) is the suggestion from Fig. 2 that for the more severe health states, TTOP values are lower than VAS ones (notwithstanding the non-monotonicity at very low values discussed above). Given that this is not the case for any of the other methods, it is likely to be explained in terms of the differential effect that TTOP has on valuations. One possible explanation is that TTOP (which has a 10-year scale on a board) is the method which focuses the respondent's attention explicitly on the length of time spent in a health state. If the disutility associated with a severe health state increases as the time spent in that state increases (as postulated by Sutherland *et al.* [18]), then lower valuations will be elicited for such states from methods which focus more explicitly on the time dimension.

Of course, if this is the reason that TTOP valuations are lower than VAS ones, then one would expect to observe a similar (although perhaps less powerful) effect with respect to TTONP, which (unlike both SG variants) also asks respondents to think in terms of time. That this is not the case raises the more general issue of why valuations from the No Props variants are higher than those from the Props variants, particularly for lower VAS values. A possible explanation relates to 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. This may have resulted in an analogue of the reference point effect, in which respondents gave special status to favourable outcomes and hence to higher (inferred) health state valuations (as evidenced by the No Props mapping functions which are above the 45° line and "fan in" from above). This suggests that No Props variants may introduce systematic bias into valuations.

It was also hypothesised that mapping functions for the SG would be different from those for the TTO; specifically, that, for the same VAS values, SG values would be higher than TTO ones. In terms of the relationship between the different choice-based methods, two patterns emerge from the mapping functions: (1) for VAS scores above about 0.4, SGP values are lower than TTOP ones, whilst SGNP and TTONP values are very similar; and (2) for VAS scores below 0.4, SGP values are higher than TTOP values and SGNP values are higher than TTONP ones.

That TTOP valuations are higher than SGP valuations for high VAS values goes against *a priori* expectations, but might be explained in terms of the relative weight respondents attach to the numeraire they are asked to sacrifice in order to gain an improvement in health. In the SG, health improvements are valued in terms of the level of risk (usually of immediate death) a respondent is prepared to accept, whilst in the TTO they are valued in terms of the amount of life expectancy a respondent is prepared to sacrifice. Thus, the results may indicate that sacrificing an extra six months of life expectancy is more valuable to respondents *in this study* than taking an extra 5% risk of death. That the expected relationship between SG and TTO holds for lower VAS values might be explained in terms of the explicit reference to the time spent in the health state in the TTO exercise.

With respect to the impact of respondent background characteristics, it is found that only age appears to have a significant impact on the resultant mapping functions. For both SG variants, it was found that those aged 18–29 years had significantly lower intercepts and steeper slopes than older respondents, suggesting that, for the same (low) VAS score, younger respondents tend to give lower valuations on the SG, but that this difference decreases as the VAS score increases. Although the same pattern is observed for the TTO variants, the

coefficients on the relevant variables (YOUNG and YOUNG*VAS) fail to reach conventional levels of significance.

It is unclear why younger respondents should have different mapping functions from other respondents, or why this difference should be more pronounced at the lower end, or why it should be more pronounced on the SG than on the TTO. The literature to date does not help to shed much light on this subject since, although age is generally regarded as having a negligible impact on health state valuations [15], much of the analysis has concentrated on differences *within* valuation methods and not *across* them.

Nonetheless, the results suggest that, at least when considering severe states of health, younger respondents (for the same VAS score) are more willing to sacrifice life expectancy and, relatively speaking, even more willing to risk death than are older respondents. It might be that younger respondents differ from older respondents more in their attitude to risk (as measured by different SG values) than in their attitudes towards time (as measured by different TTO values). However, that both these differences decrease as the VAS value increases suggests that either risk attitude and time preference are not independent of the health state being valued or that something else is being picked up here. These would appear to be important issues that future (quantitative and qualitative) research efforts should be directed towards addressing.

This paper has attempted to assess whether VAS valuations can be mapped into SG and/or TTO ones. If robust mapping functions could be estimated, then this would have the practical advantage of allowing valuations elicited from the cheap and simple VAS to be converted into theoretically superior SG and/or TTO ones. In addition, models could be developed that might explain the mapping functions, in the same way that Loomes [19] has shown that Regret Theory in its non-stochastic form can explain the relationship between aggregate values in the Torrance [5] and Wolfson *et al.* [6] data.

However, the results presented in this paper suggest that the way the methods are administered is as important a determinant of the resultant mapping functions as the methods are themselves. This suggests that the way in which a question is framed can have a significant effect on responses, a fact which is increasingly recognised by many economists who now accept that changes in questionnaire design can bias a respondent's stated preferences. This may be because the different designs tap into different parts of a respondent's preference function, or may be because they simply confuse respondents and hence confound their preferences.

More fundamentally, the importance of framing effects may call into question the view held by most economists that people have clear, well-defined preferences (referred to by Fischoff [20] as the "theory of articulated values")—at least with respect to health. Whatever the *cause* of the importance of questionnaire design (a detailed discussion of which is beyond the scope of the present paper), the results in this paper suggest that one important *consequence* is that no single set of mapping functions is likely to explain the observed disparities between health state valuation methods.

Acknowledgements—This paper reports on data collected as part of a wider study conducted by the Measurement and Valuation of Health (MVH) Group at the Centre for Health Economics, University of York, in collaboration with Social and Community Planning Research. The study and the MVH Group (of which Paul Dolan was a member) were financed by the Department of Health. The comments of two anonymous referees are gratefully acknowledged.

REFERENCES

1. Berzon, R. and Shumaker, S. (1993) A critical review of cross national health-related quality of life instruments. *Quality of Life Newsletter* **5**, 1–2.
2. Torrance, G. W. (1986) Measurement of health state utilities for economic appraisal. *Journal of Health Economics* **5**, 1–30.
3. Richardson, J. (1994) Cost–utility analysis: what should be measured? *Social Science & Medicine* **39**(1), 7–21.
4. Gafni, A. and Torrance, G. W. (1984) Risk attitude and time preference. *Management Science* **30**, 440–451.
5. Torrance, G. W. (1976) Social preferences for health states: an empirical evaluation of three measurement techniques. *Socio-economic Planning Sciences* **10**, 129–136.
6. Wolfson, A. D., Sinclair, A. J., Bombardier, C. and McGeer, A. (1982) Preference measurements for functional status in stroke patients: inter-rater and inter-technique comparisons. In *Values and Long Term Care*, eds R. L. Kane and R. A. Kane, pp. 191–211. Lexicon, MA.
7. Read, J. L., Quinn, R. J., Berwick, D. M., Fineberg, H. V. and Weinstein, M. C. (1984) Preferences for health outcomes: comparison of assessment methods. *Medical Decision Making* **4**, 315–329.
8. Hornberger, J. C., Redelmeier, D. A. and Petersen, J. (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.
9. Morris, J. and Durand, A. (1989) Category rating methods: numerical and verbal scales. Mimeo, Centre for Health Economics, University of York, York.
10. Kahneman, D. and Tversky, A. (1979) Prospect theory: an analysis of decision under risk. *Econometrica* **47**(2), 263–291.
11. EuroQol Group (1990) EuroQol: a new facility for measurement of health-related quality of life. *Health Policy* **16**, 199–208.
12. Gudex, C. (1994) *Standard Gamble User Manual: Props and Self-Completion Method.* Occasional Paper, Centre for Health Economics, University of York, York.
13. Gudex, C. (1994) *Time Trade-off User Manual: Props and Self-Completion Method.* Occasional Paper, Centre for Health Economics, University of York, York.
14. Patrick, D. L., Starks, H. E., Cain, K. C., Uhlmann, R. F. and Pearlman, R. A. (1994) Measuring prefer-

ences for health states worse than death. *Medical Decision Making* **14**, 9–18.
15. Froberg, D. G. and Kane, R. L. (1989) Methodologies for measuring health state preferences 3: population and context effects. *Journal of Clinical Epidemiology* **42**(6), 585–592.
16. Greene, W. H. (1992) *LIMDEP Version 6.0: User's Manual and Reference Guide*. Econometric Software, New York.
17. Nord, E. (1992) Methods for quality adjustment of life years. *Social Science & Medicine* **34**, 559–569.
18. Sutherland, H. J., Llewellyn-Thomas, H., Boyd, N. F. and Till, J. E. (1982) Attitude toward quality of survival: the concept of maximal endurable time. *Medical Decision Making* **2**, 299–309.
19. Loomes, G. (1993) Disparities between health state measures: is there a rational explanation? In *The Economics of Rationality*, ed. W. Gerrard. Routledge, London.
20. Fischoff, B. (1991) Value elicitation: is there anything in there? *American Psychologist* **46**(8), 835–847.