OUTCOME MEASUREMENT

THE TIME TRADE-OFF METHOD: RESULTS FROM A GENERAL POPULATION STUDY

PAUL DOLAN[1], CLAIRE GUDEX[2], PAUL KIND[2] AND ALAN WILLIAMS[2]
[1] *University of Newcastle, Newcastle-Upon-Tyne, UK;*
[2] *Centre for Health Economics, University of York, UK.*

SUMMARY

An important consideration when establishing priorities in health care is the likely effects that alternative allocations of resources will have on health-related quality-of-life (HRQoL). This paper reports on a large-scale national study that elicited the relative valuations attached by the general public to different states of health (defined in HRQoL terms). Health state valuations were derived using the time trade-off (TTO) method. The data from 3395 respondents were highly consistent, suggesting that it is feasible to use the TTO method to elicit valuations from the general public. The paper shows that valuations for severe health states appear to be affected by the age and the sex of the respondent; those aged 18–59 have higher valuations than those aged 60 or over and men have higher valuations than women. These results contradict those reported elsewhere and suggest that the small samples used in other studies may be concealing real differences that exist between population sub-groups. This has important implications for public policy decisions.

KEY WORDS—health status measurement; time trade-off; utility.

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 benefits of the alternative uses of scarce resources should be taken into account. An important part of the benefit of any health care intervention is its effect on the health-related quality-of-life (HRQoL) of the population it affects, which will ultimately be the general public given that they are all potential patients. Of course, the views of the general public will also be relevant in their capacity as taxpayers.

From the preferences of the general public, a set of values for the whole community can be built up. Also, any sub-groups that have markedly different valuations from the rest can be identified, and a separate set of valuations calculated for them. This information could then be used in a variety of ways; for example, in clinical trials where HRQoL is an important feature, in association with population surveys to measure levels and trends in community health, and in the calculation of Quality-Adjusted Life-Years (QALYs).

Thus, the objective of this study was to establish the relative valuations attached to different states of health (defined in HRQoL terms) by members of the general public.

Health state valuations can be elicited by using a number of different methods.[1,2] Economists have tended to prefer the standard gamble (SG) method, because of its foundations in von Neumann-Morgernstern Expected Utility Theory. However, the time trade-off (TTO) has also been widely used, and shares a common theoretical foundation with the SG in utility theory generally i.e. they both require people to sacrifice one thing they value

Address for correspondence: Paul Dolan, Department of Economics, University of Newcastle, Newcastle-Upon-Tyne NE1 7RU, England.

CCC 1057–9230/96/020141–14

Received 22 September 1994
Accepted 9 January 1996

(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,[3,4] a choice between them needs to be informed by their respective performance on empirical grounds.

A recent (within-respondent) comparison of the two methods, suggested that the TTO performed slightly better in terms of the internal consistency of the answers given by respondents, the sensitivity of valuations to parameters known to influence them, and the reliability of the responses when the valuation task is repeated by the same respondents some weeks later.[5] Thus, it was decided to use the TTO in this study.

STUDY DESIGN

Sample selection

In determining the size of the sample, there was the need for enough observations to be obtained so as to detect differences between the valuations given to different states and to be able to detect differences in valuations between different subgroups of the population (e.g. by age, or social class, or geographical location). Although there is little evidence in the literature regarding what size difference is required to be considered meaningful,[6] it was decided that a 0.05 difference between health states and between different subgroups is likely to be considered important in many contexts. A sample size of 3235 enabled such a difference to be detected between health states and between four equally-sized subgroups

Mobility
1. No problems in walking about
2. Some problems in 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

Note: 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 in 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

Figure 1. The EuroQol Descriptive System

at the 0.05 level of significance with 80% power. This required the selection of 6080 addresses; thus allowing for a response rate of 53%. The sample was drawn up by Social and Community Planning Research (SCPR) using the postcode address file. The main fieldwork was carried out by 92 trained interviewers between August and December 1993.

Choice of health states

Health states were defined in terms of the Euroqol Descriptive System (see Fig. 1) which generates $3^5=243$ theoretically possible health states. (For completeness, two additional 'states' —'Unconscious' and 'Immediate Death'—are added, yielding 245 states in total.) For the EuroQol to be used in evaluating the health benefits associated with different health care interventions, it is important to derive a single index value for each of these states. (Of course, some of these states, for example, 33331, might be considered to be highly implausible but circumstances where they exist cannot be ruled out *ex ante*). Previous piloting showed no one respondent can be expected to value more than about 13 states using the TTO in any one interview but this number was deemed to be too small to interpolate valuations for all possible EuroQol states from. Therefore, a larger set of 43 states was chosen in total and each respondent was asked to value a subset of these.

In choosing the states both for use in the study itself and for each respondent, the most important consideration was that they should be widely spread over the valuation space so as to include as many combinations of levels across the five dimensions as possible. This was subject to the constraint that the states were likely to be considered plausible by respondents. Therefore, level 1 on usual activities (no problems) was not combined with level 3 on mobility (confined to bed) or with level 3 on self-care (unable to wash or dress self). Figure 2 shows the set of states chosen for direct valuation and how a subset of these were chosen for each respondent.

Structure of the interview

Each respondent was first asked to describe their own health using the EuroQol descriptive system. They were then asked to rank a predetermined set of 15 health states (the 13 to be used in the TTO plus 11111 and 'Immediate Death'), which were printed on cards, in order from best to worst. It was explained that each state was to be regarded as lasting for 10 years without change, followed by death. The respondent was then asked to indicate where on a vertical VAS with endpoints of 100 (best imaginable health state) and 0 (worst imaginable health state) they would rate each of the states.

The 13 states were then valued by the TTO method using a specially-designed double-sided board. One side was relevant for states that were

Each respondent valued 11111, Immediate Death, 33333 and unconscious

plus

2 from 5 'very mild' states:

11112 11121 11211 12111 21111

plus

3 from 12 'mild' states:

11122 11131 11113 21133 21222 21312 12211 11133 22121 12121 22112 11312

plus

3 from 12 'moderate' states:

13212 32331 13311 22122 12222 21323 32211 12223 22331 21232 32313 22222

plus

3 from 12 'severe' states:

33232 23232 23321 13332 22233 22323 32223 32232 33321 33323 23313 33212

Figure 2. Health states valued in the study

regarded by the respondent as better than dead, and the other side for states that were regarded as worse than dead. In the former case, respondents were led by a process of 'bracketing' to select a length of time in the 11111 state that they regarded as equivalent to 10 years in the target state; the shorter the 'equivalent' length of time, the worse the target state. Respondents were given an opportunity to refuse to trade-off any length of life in order to improve its quality. In the case of states worse than dead, the choice was between dying immediately and spending a length of time (x) in the target state followed by (10 - x) years in the 11111 state; the more time required in the 11111 state to compensate for a shorter time in the target state, the worse the target state. For further details of the protocol used see.[7] At the end of the interview, personal background data were collected from each respondent.

Ten years was chosen as the time horizon because it was considered long enough for respondents to be able to make meaningful sacrifices and to be able to distinguish between states but not too long so as to be unrealistic for older respondents. It is recognised that this time horizon would have been unrealistically short for many younger respondents but it was felt that other alternatives (such as variable time horizons based on a person's own expected life expectancy) would have created even greater problems of measurement and interpretation.

Retest interview

In order to test the reliability of the TTO valuations, a sub-sample of 221 respondents that were representative of the full sample in terms of sex, age, and qualifications were taken through exactly the same interview by the same interviewer about 10 weeks after the original interview.

STUDY POPULATION AND EXCLUSIONS

Of the 6080 addresses selected for sampling, 706 (12%) were found to be 'out of scope', being non-

Table 1. Characteristics of the sample (Figures and percentages)

Characteristic	Full Sample (*n* = 3395)	After Exclusions (*n* = 3337)	GHS
Sex: Male	43	43	47
Female	57	57	53
Age: 18–34	31	32	31
35–49	25	25	27
50–59	14	14	15
60+	31	30	28
Education: Degree	9	9	8
Higher	11	11	10
A/O levels	40	41	45
None	37	37	35
Foreign/Other	3	3	3
Social Class: I, II	29	30	30
III Non-manual	24	24	22
III Manual	20	21	21
IV, V	25	25	21
Other	1	1	3
Marital status: single	17	17	21
married	60	60	64
widowed	13	12	9
divorced	10	11	6
Those reporting problems on:			
Mobility	18.4	18.1	—
Self-care	4.2	4.2	—
Usual activities	16.3	16.2	—
Pain/discomfort	32.9	32.8	—
Anxiety/depression	20.9	20.8	—

residential, empty/derelict, untraceable, or not yet built. Of the remaining 5324 addresses, 3395 interviews were achieved, giving a response rate of 64% on in-scope addresses. The main reasons for unsuccessful interviews were a refusal by the selected person. Table 1 shows that the sample had broadly similar characteristics in terms of age, sex, marital status. educational attainment and social class as the general population. Table 1 also shows the number (and background characteristics) of respondents excluded from subsequent data analysis. Because the criteria for excluding respondents were as stringent as possible, in total only 58 (1.3%) of respondents were excluded: 42 had insufficient data for further analysis; 7 had rated all states as worse than death; and 9 did not

Table 2. TTO valuations (when scores range from 1 to −1)

State	*N*	Mean (SD)		Median (IQR)	
21111	1306	0.87	(0.24)	0.95	(0.83 - 1.00)
11211	1335	0.87	(0.23)	0.95	(0.83 - 1.00)
11121	1310	0.85	(0.25)	0.93	(0.80 - 1.00)
12111	1310	0.83	(0.30)	0.93	(0.80 - 1.00)
11112	1309	0.82	(0.29)	0.93	(0.75 - 1.00)
12211	828	0.76	(0.33)	0.90	(0.63 - 1.00)
12121	828	0.74	(0.32)	0.85	(0.60 - 1.00)
11122	816	0.72	(0.37)	0.83	(0.63 - 1.00)
22121	830	0.64	(0.42)	0.78	(0.50 - 0.93)
22112	840	0.66	(0.38)	0.74	(0.50 - 0.95)
11312	824	0.55	(0.47)	0.68	(0.40 - 0.93)
21222	823	0.55	(0.46)	0.65	(0.40 - 0.91)
12222	830	0.54	(0.47)	0.65	(0.38 - 0.93)
21312	811	0.51	(0.49)	0.65	(0.33 - 0.93)
22122	809	0.53	(0.47)	0.63	(0.39 - 0.93)
22222	834	0.50	(0.49)	0.63	(0.35 - 0.88)
11113	823	0.39	(0.56)	0.50	(0.00 - 0.88)
13212	820	0.38	(0.54)	0.50	(0.04 - 0.78)
13311	810	0.33	(0.56)	0.50	(0.00 - 0.75)
11131	812	0.20	(0.60)	0.38	(−0.33 - 0.72)
12223	828	0.21	(0.56)	0.35	(−0.28 - 0.63)
21323	819	0.15	(0.59)	0.30	(−0.38 - 0.60)
23321	821	0.14	(0.61)	0.30	(−0.41 - 0.63)
32211	833	0.14	(0.60)	0.25	(−0.38 - 0.63)
21232	826	0.06	(0.61)	0.13	(−0.48 - 0.55)
22323	812	0.04	(0.59)	0.03	(−0.48 - 0.53)
33212	829	−0.02	(0.60)	0.00	(−0.50 - 0.48)
23313	830	−0.07	(0.58)	0.00	(−0.55 - 0.40)
22331	814	−0.01	(0.60)	0.00	(−0.53 - 0.50)
11133	829	−0.05	(0.61)	0.00	(−0.58 - 0.48)
21133	826	−0.07	(0.59)	−0.03	(−0.60 - 0.45)
23232	827	−0.10	(0.59)	−0.08	(−0.63 - 0.43)
33321	828	−0.14	(0.57)	−0.23	(−0.63 - 0.38)
32313	832	−0.16	(0.57)	−0.23	(−0.63 - 0.30)
22233	829	−0.15	(0.57)	−0.28	(−0.63 - 0.34)
32223	825	−0.19	(0.56)	−0.28	(−0.68 - 0.23)
13332	812	−0.23	(0.55)	−0.38	(−0.70 - 0.18)
32232	818	−0.23	(0.57)	−0.38	(−0.73 - 0.20)
32331	826	−0.27	(0.55)	−0.38	(−0.78 - 0.03)
Uncon	3294	−0.41	(0.39)	−0.38	(−0.83 - −0.03)
33232	824	−0.33	(0.51)	−0.43	(−0.75 - 0.00)
33323	833	−0.39	(0.49)	−0.48	(−0.83 - −0.03)
33333	3289	−0.54	(0.41)	−0.65	(−0.93 - −0.28)

understand the TTO task. It can be seen that excluded respondents were more likely to be aged 60 and over, to have no qualifications, to be in social classes III-V and to report problems on the EuroQol dimensions. However, given such a small number of exclusions, the 3337 respondents remaining in the data set were still broadly representative of the general population.

VALUATION RESULTS

Adjustment of scores

If full health and dead are assigned scores of 1 and 0 respectively, then for states that are rated as better than dead on the TTO, scores are given by the formula x/10 where x is the number of years spent in full health. For states that are rated as worse than dead, the score is given by the formula $-x/(10-x)$ i.e. negative scores lie on a ratio (not an interval) scale. However, Eyman[8] (Table 7), demonstrated that values generated in this way can lead to biases in observers' judgements. Poulton[9] describes how such biased judgements can be corrected. The implication of his work, applied to states worse than dead, is that these values should be treated as having interval scale properties. Hence these valuations have been transformed using the formula $(x/10)-1$. Scores for states worse than dead are now bounded by -1, just as states which are better than dead are limited by a value of 1 for full health. This transformation is used elsewhere in the literature.[10]

Distribution of scores

Table 2 shows the transformed mean and median scores for all 43 states. Inspection of the range of health state values suggests that respondents were more prepared to sacrifice life expectancy for states that include 'extreme problems' with any of the dimensions. Level 2 (which involves 'some problems') on the dimensions appears to be much more tolerable. For example, state 22222 has a median valuation that is 0.13 higher than 11113 and 0.25 higher than 11131. This results in most states that include level 3 on two or more dimensions having values that imply they are, on average, perceived to be worse than death. Taking mean scores, 17 states had a negative score and there were 13 states with a negative median score, and a further 4 had median values of 0.0 (i.e. they were rated as bad as being dead).

Kolmogorov-Smirnov tests indicated that the distribution of scores for each state was non-normal: the distributions were generally negatively-skewed for less severe states (indicated by higher median than mean values for such states) and positively-skewed for more severe states (as evidenced by higher mean values for such states). Because the data was not readily transformed to a more sensible (normal) distribution, non-parametric Mann-Whitney U tests and Wilcoxon matched-pairs

Table 3. Definition of variables used in regression analysis

Variable	Definition
SEX	A dummy taking the value of 1 if the respondent is female, and 0 otherwise
AGE	A continuous variable for the respondent's age
AGE2	Age-squared
EDU1	A dummy taking the value of 1 if the respondent has intermediate qualifications, and 0 otherwise
EDU2	A dummy taking the value of 1 if the respondent has no qualifications, and 0 otherwise
SOC1	A dummy taking the value of 1 if the respondent is in social class III, and 0 otherwise
SOC2	A dummy taking the value of 1 if the respondent is in social class IV or V, and 0 otherwise
MAR1	A dummy taking the value of 1 if the respondent is separated, divorced or widowed, and 0 otherwise
MAR2	A dummy taking the value of 1 if the respondent is single, and 0 otherwise
MOB	1 if the respondent reports problems on mobility, 0 otherwise
SELF	1 if the respondent reports problems on self-care, 0 otherwise
UACT	1 if the respondent reports problems on usual activities, 0 otherwise
PAIN	1 if the respondent reports pain/discomfort, 0 otherwise
MOOD	1 if the respondent reports anxiety/ depression, 0 otherwise

signed-rank tests (as opposed to parametric t-tests) have been used where appropriate.

The effect of background characteristics

Before addressing this issue, it was determined that the valuations were not susceptible to interviewer bias nor to regional effects. Ordinary least-squares regression analysis was used to assess the impact of a number of respondent characteristics on health state valuations. The dependent variable was taken to be the TTO valuation and the independent variables were the different background characteristics (see Table 3 for a description of these variables). Most of the variables are categorical except for age which is a continuous variable and age-squared which of the various transformations of age tested was found to be the most significant. To allow for the possibility that the impact of one or more of these variables may not be uniform across the entire range of EuroQol states, the regression was performed separately on the 'mild' (which included the set of five 'very mild' states), 'moderate' and 'severe' states, as defined in Figure 2.

Table 4. Results of regression analysis

Variable	Mild states	Moderate states	Severe states
Adjusted R^2	0.01	0.02	0.03
Sample	16459	9880	16470
Constant	0.511 (14.495)	0.118 (2.129)	-0.262 (-6.808)
SEX	−0.029 (−3.655)	−0.063 (−5.008)	−0.070 (−7.987)
AGE	0.008 (5.839)	0.010 (4.815)	0.008 (5.400)
AGE2	−0.00010 (−7.114)	−0.00013 (−6.458)	−0.00012 (−8.100)
EDUL	0.003 (−1.804)	−0.001 (−0.377)	−0.004 (−2.432)
EDU2	0.0004 (0.370)	−0.001 (−0.400)	−0.001 (−1.777)
SOC1	−0.003 (−2.146)	0.002 (1.037)	−0.0003 (−0.242)
SOC2	0.0003 (0.305)	−0.001 (−0.544)	−0.003 (−2.992)
MAR1	−0.006 (−4.347)	−0.007 (−3.855)	−0.003 (−2.320)
MAR2	−0.005 (−4.324)	−0.004 (−2.627)	−0.003 (−2.114)
MOB	0.037 (3.095)	0.064 (3.401)	0.039 (3.099)
SELF	0.001 (0.132)	0.008 (0.469)	0.010 (0.880)
UACT	0.007 (0.675)	−0.004 (−0.251)	0.013 (1.079)
PAIN	−0.0003 (−0.037)	−0.044 (−3.166)	−0.049 (−5.068)
MOOD	−0.011 (−1.142)	0.011 (0.739)	0.007 (0.719)

The results are shown in Table 4. That the adjusted- R^{2}s are so low is not in itself a cause for much concern since the object of this analysis is to assess the *relative* effect of different respondent characteristics on valuations rather than to find the model(s) which explains *all* the variance in valuations. Given the large number of observations in each regression, a particular variable is considered to be significant if the (absolute) t-statistic associated with it is greater than 3.29 (which corresponds to a probability value of 0.001).

The results suggest that TTO valuations are primarily affected by the age and sex of the respondent. Figure 3 shows the effect of age on the valuations given by men and women, respectively, when all other dummies take a value of zero. They suggest that TTO valuations increase slowly from the age of 18 to about 40, then begin to fall slowly from about 40 to 60 and then fall sharply in later years. Although this pattern is observed for all three sets of states, it is more marked for moderate and severe states than for the mild states. The effect of gender is also more pronounced for more severe states: for the set of mild states, women give valuations that are, on average, 0.03 lower than those given by men but the difference increases to 0.06 for moderate states and to 0.07 for severe states.

In addition to age and sex, the marital status of the respondent appears to be a statistically significant explanatory variable for the set of mild states (where, on average, the valuations of single people are 0.006 higher than the valuations of married people and 0.005 lower than those who are separated, divorced or widowed) and for the set of moderate states (where single people have valuations that are 0.008 lower than married people). However, it can be seen that the value of these

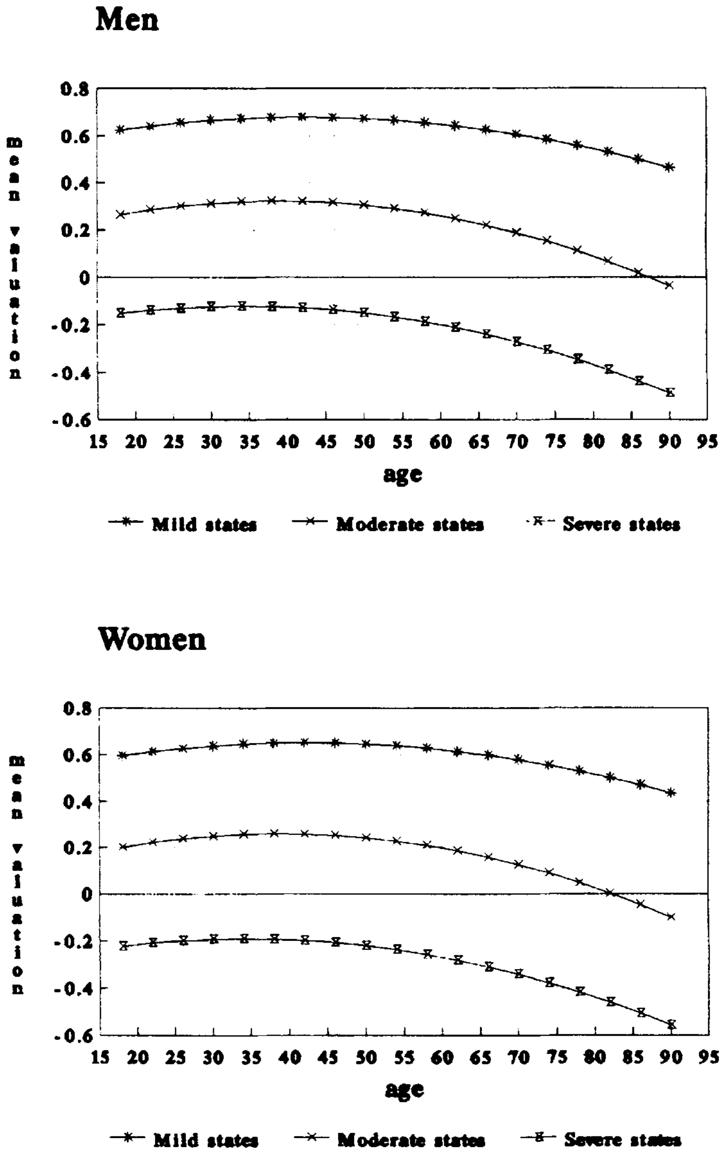

Figure 3. The effect of age on valuation

coefficients is very small suggesting that, although statistically significant in the regression equation, the effect of marital status is negligible and unlikely to be meaningful in any practical sense. Finally, a respondent's description of their current health status on the EuroQol dimensions is seen to have some effect on valuations although the direction of this effect is not systematic across dimensions or states: for the set of moderate states, those with problems walking about have *higher* valuations than those with no problems walking about, and, for the set of severe states, those reporting pain have *lower* valuations than those reporting no pain.

Quality versus quantity?

The question of whether TTO valuations differ by sub-group is essentially about whether some people are more or less prepared to sacrifice life expectancy in order to avoid poor health than other people. In this context, there is another important question; namely, are some people more or less willing to sacrifice *any* life expectancy in order to avoid poor health than others? This draws a distinction between those willing to trade quantity (in terms of life expectancy) for quality (in terms of improvements in health), irrespective of the rate of exchange, and those unwilling to 'play the game'. In other words, there exists a qualitative difference between an implied health state value of 1.00 and any other value.

46% of respondents were willing to sacrifice life expectancy to avoid *all* of the dysfunctional states they were presented with, and thus had no health state values of 1.00. A further 29% were willing to sacrifice life expectancy for all but one or two of the states. In such cases, the unwillingness to trade-off time was almost exclusively associated with one or both of the very mild states. In all, 95% of respondents were prepared to sacrifice life expectancy for 6 or more states. The 25% of respondents who were unwilling to trade off time for three or more states were older and less educated than the remainder of respondents; 33.7% were aged 60 or over compared to 28.4% and 41.8% had no qualifications compared to 34.7% in the group of respondents more willing to sacrifice life expectancy. Interestingly, the 5% of respondents who were unwilling to sacrifice any life expectancy in order to avoid more than half of the states they valued were no older than the remainder of respondents. Instead, such respondents were found to be less educated (45.6% had no qualifications compared with 33.9% of the other respondents).

TEST-RETEST RELIABILITY

The 221 respondents in the retest were representative of those in the test in all respects except educational level, where 28.6% of retest respondents had no qualifications compared with 37.0% of respondents *not* in the retest (Chi = 6.26, d.f. = 1, $p < 0.05$). For the purposes of group analysis, 4 respondents were excluded from the retest data set: 1 previously excluded from the test data set; 1 with all states missing at retest; 1 with all states rated as worse than dead; and 1 with the same score given to all states. At test, respondents taking part in the retest gave a significantly higher TTO score to state 33323 than respondents who did not go on to do the retest ($p < 0.01$), but this was the only significant difference in the valuations given by the two groups of respondents. The results of Wilcoxon matched-pairs signed-rank tests showed that no health state valuation at retest was significantly different from its corresponding value at test. However, Figure 4, which graphically represents the differences in median scores between test and retest, shows that for 2 (3) states the difference between the median at test is more than 0.20 higher (lower) than the median at retest.

For comparisons on an individual-by-individual basis an intra-class correlation coefficient (ICC) was calculated for each respondent for each of the valuation methods. This statistic is calculated using the following formula;

$$\mathrm{ICC} = (A^2 + B^2 - C^2)/(A^2 + B^2 + D^2 - C^2)$$

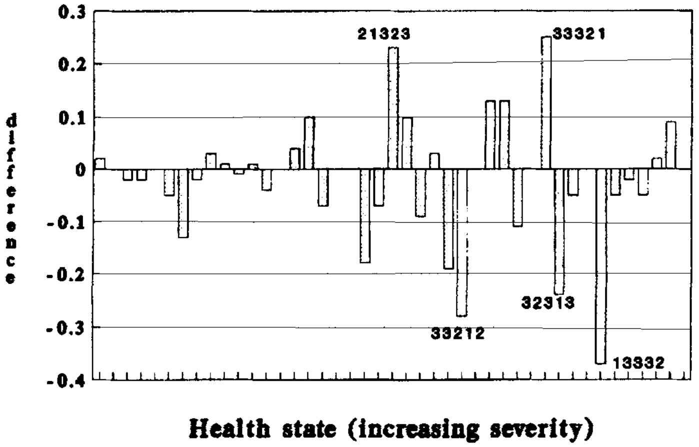

Figure 4. Difference between test and retest: median at test minus median at retest

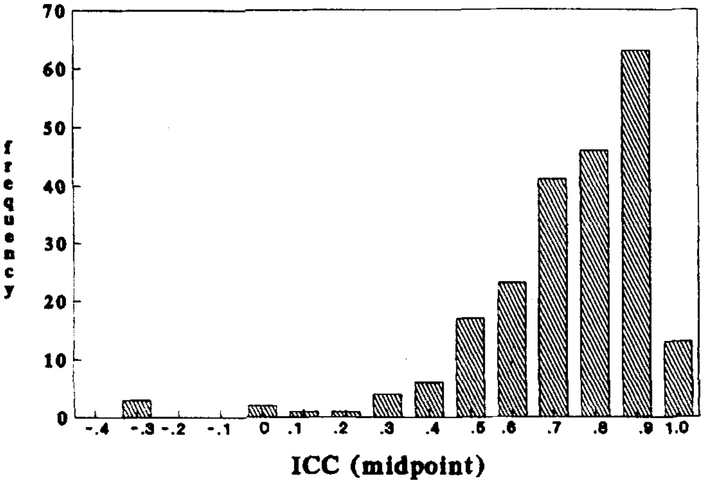

Figure 5. Distribution of ICCs

where

A is the SD of the difference between each score at test and the mean score at test

B is the SD of the difference between each score at retest and the mean score at retest

C is the SD of the difference between each score at test and each score at retest

D is the mean difference between each score at test and each score at retest

The closer the ICC is to 1, the greater the reliability. Figure 5 shows the distribution of ICCs. The majority of respondents had an ICC that was close to 1 and only 24 (10.9%) had an ICC that was less than or equal to 0.5. The mean ICC was 0.73 (S.D. = 0.22) and the median was 0.79 (IQR = 0.64 – 0.88).

ICCs appeared to be negatively related to educational attainment; those with a degree or equivalent had higher ICCs as a group than those with no qualifications at all ($p < 0.05$).

DISCUSSION

The group valuations elicited for the 43 health states suggest that members of the general public can distinguish between states of health that involve different degrees of severity. However, the measures of dispersion (SDs and IQRs) were much higher than expected which casts doubt on the assertion made by Torrance[1] that 'the mean utility value for a health state can be made as precise as desired by increasing the group size'. Rather than reflecting the degree of consensus about the value that should be attached to a particular health state, it is possible that the large SDs and IQRs reflect the difficulties respondents encountered in imagining themselves being in the health states so described. That the variance around the central tendency values increases as the severity of the health state increases, lends some support to this hypothesis.

However, the interpretation of measures of dispersion does not tell the whole story, because it is quite plausible that respondents rank adjacent states in the same way, but some do so using high values, while others do so using low values. Analysing pairwise relationships between states (using the Wilcoxon matched-pairs signed-rank test), revealed that there were no more than 4 states adjacent to any particular state which were *not* significantly different from it at the 1% level. Thus, it appears that the large SDs and IQRs obtained in this study, particularly for the more severe health states, are more likely a reflection of the fact that different people have very different views about the same health state, rather than an indication of respondent confusion.

It is unclear how generalizable these results are since it is likely that they are in part a function of the duration of the states. Since respondents were told to imagine that each state would last for ten years *without any change*, it is likely that some felt they could not tolerate extreme dysfunction (particularly pain) for this long. Whilst the finding that some states were considered worse than death is not unique (they have appeared in several countries for several valuation methods,[11,12]) there is evidence to suggest that fewer states would be regarded as worse than death were they to last for less time (see Sutherland *et al*[13] who postulate the concept of a 'maximal endurable time', after which some states yield a negative utility). There is the general question of how valuations for health states are affected by duration when using *any* of the valuation methods, and it is one which requires urgent attention.

Related to this, is the impact of time preference on valuations elicited by the TTO method. For TTO valuations to be interpreted as index values between −1 and 1, requires each year of life to be valued equally. However, if people discount future years of life because of a positive rate of time preference (i.e. because they give greater value to years of life in the near future than to those in the distant future), then it is no longer valid to treat TTO valuations in this way. Moreover, if people are not prepared to trade-off a constant proportion of their remaining life expectancy in order to avoid a dysfunctional health state, then valuations elicited for states lasting ten years cannot be assumed to hold for states lasting for longer or shorter durations *irrespective of the impact of duration*. These issues are yet be resolved satisfactorily in the literature.

There is also the issue of whether the order of presentation for states rated as worse than death may have had an effect on valuations: respondents may value a scenario in which a bad state is followed by a good state (as in this study) differently from one in which a good state is followed by a bad state (as suggested by Torrance[1]), even though the time spent in each of the states may be identical. This is an empirical question which needs addressing. In addition, there is the question of how to interpret scores for states worse than dead. Given the standard health preference scale, states preferred to death have an upper bound of one but there is no comparable lower bound for states rated worse than death. The asymmetry results from the TTO (as well as for the standard gamble) producing an interval scale for positive scores and a ratio scale for negative scores. It seems reasonable to treat positive and negative scores in the same way i.e. to convert the ratio scores into interval ones, thus setting a lower bound of −1, and this adjustment finds support in the psychometric literature.[9]

One of the most important findings is the effect that the age of the respondent had on health state valuations. The valuations of 'middle-aged' respondents appear to be higher than those of younger respondents, whilst older respondents have much lower valuations than those in the other two age 'groups'. This may lend support to the notion that the middle-aged have the lowest rates of time preference and thus place relatively more weight on years in the future (i.e. the ones they are being asked to sacrifice) than younger or older respondents. However, the fact that the effect of age is not uniform across all states, being more pronounced for moderate and severe states than for mild ones, may suggest that the effect of time preference for health is *not* independent of the severity of the health state.

It may be that the much lower valuations of the older respondents in this study are an artefact of the TTO method. For states that were rated as better than dead, respondents were asked to imagine that each state would last for 10 years without any change, after which they would die. If they did not believe that they actually had 10 years life expectancy, they might willingly give up these 'excess' life years, thereby depressing the apparent value attached to the health states. However, the effect of age appears to be more pronounced for the more severe states (which were much more likely to be rated as worse than dead) than for the less severe ones. It is unclear how and why an argument of this kind would apply with greater force to the worse than dead scenario than to the better than dead one.

An alternative explanation is that, as people's life expectancy shortens, they see less reason to tolerate suffering during their remaining years. Conventional wisdom suggests that people become *more* tolerant of poor health as they get older, either through adapting to a general deterioration in health or through a lowering of expectations, and there is some empirical evidence to support this hypothesis.[14] However, it is entirely plausible that somebody who has limited life expectancy and is possibly in a poor health state, may be prepared to sacrifice a great deal (either life expectancy for states rated better than dead or time in full health for states rated worse than dead) in order to avoid severe health states. In a study of cancer and renal patients with limited life expectancy, Shiell, King and Briggs [15] found that TTO results were polarised; some would not trade off any life years, while others would trade off almost everything to have their final years as healthy ones. The older respondents in this general population study may have held similar views about the (severe) states as this latter group.

In addition, older respondents may be more conscious of the burden that serious chronic illness can place on their family or close friends, particularly if they have experienced the suffering of someone close to them. This might explain why the valuations of older respondents were closer to those of other respondents for states they considered 'tolerable' (both for themselves and for those close to them) yet much lower for states they considered would be 'intolerable' for themselves and their family. It may also go some way towards explaining why women had lower valuations than men for the more severe states: women may be more concerned about the burden they would be to others than men are, particularly as they may be likely to have experience of caring for someone with serious illness.

The premise of the discussion so far has been that differences in valuations according to the age and, to a lesser extent, gender of the respondent are *real* differences. However, it could be argued that this relationship is a spurious one; that the large number of variables being assessed and the large sample size, by chance, account for the results. Whilst this possibility cannot be completely ruled out, the fact that in the three regressions (one for each set of mild, moderate and severe states) the effects of age and sex are systematic (though not constant) suggests that genuine effects are being picked up. In addition, the use of regression analysis should isolate the effects the age and gender, and thus reduce the possibility that they are acting as proxies for other (more important) explanatory variables.

Although almost half of the respondents were prepared to sacrifice life expectancy in order to avoid *all* of the dysfunctional states they were asked to consider, one-quarter were unwilling to sacrifice even a couple of weeks at the end of 10 years for 3 or more states. Such preference may be indicative of a bias in favour of the status quo. The notion of a status quo effect is, essentially, that people may give some special status to their current position, and react asymmetrically to movements away from that position, placing greater weight on what they perceive as losses *vis-a- vis* the status quo than on what they perceive as gains. The frequently observed substantial disparities between what people say they would be willing to pay (WTP) for some marginal benefit, and what they would be willing to accept as monetary compensation for a comparable marginal disbenefit, is often taken as evidence of such an effect. (See Kahneman, Knetsch and Thaler[16] for a general review of earlier evidence, plus fresh evidence of their own; see also Dolan, Jones-Lee and Loomes[17] for evidence of such effects in the context of road safety).

How might such effects operate on the data generated by the TTO? The answer lies in the fact that for the TTO method (unlike for the VAS, for example) respondents are asked to imagine that they are already in the poor health state. They are asked then to sacrifice some of their 'endowment' of life years in order to be in full health. If perceived losses (in terms of life expectancy) are weighted more heavily than perceived gains (in terms of HRQoL), the effect will be to elicit a higher health state valuation than would be the case otherwise. At the limit, the perceived loss is so great as to make any change undesirable. It seems reasonable to suppose that the status quo effect would be more prevalent the more unclear respondents are about the choices they are being asked to make; an 'if in doubt, stick with what you've got' hypothesis. The hypothesis might be, therefore, that less educated respondents are more likely to suffer from status quo effects. The fact that less educated respondents were more likely to be unwilling to sacrifice *any* life expectancy, even for some moderate and severe states, lends support to this hypothesis. Certainly, the possibility of a status quo effect should not be overlooked and may go some way towards explaining the observed differences between valuation methods that start with different 'endowments'.

The results from the retest were encouraging. At the aggregate level, 32 of the 43 states had a median at retest that was within 0.1 of the median at test and there were no significant differences in the valuations of any of the states between test and retest. This finding is consistent with other studies which found group values to be remarkably stable regardless of the make-up of the group.[18,19] At the individual level, only 1 in 10 of the respondents had an intra-class correlation coefficient that was below 0.5. The mean ICC was 0.73 which was deemed to be acceptable and compares well with the 0.81 found by Churchill *et al*[20] particularly when the time between test and retest was longer in this study; 10 weeks compared with 4 weeks in the Churchill *et al* study.

Of course, the stability of the valuations of the general public does not necessarily imply that the valuations of all groups will be stable. For example, this study did not include test-retest measurements taken from patients before and after therapy, whose valuations of the same state might be expected to differ. Christensen–Szalanski[21] found that women's preferences for anaesthesia during childbirth were labile; not surprisingly, perhaps, preferences for anaesthesia were more positive during labour than they were one month before or after labour. However, Llewellyn–Thomas *et al*[22] found that patients' TTO valuations of hypothetical health states encountered during radiation therapy for laryngeal cancer remained stable when those states were experienced at a later time.

CONCLUDING REMARKS

This survey was based on many years of research designed to find an effective way of generating data on health state valuations from a large representative sample of the general public. Representativeness was achieved and the data were near complete and highly consistent. The TTO data refute the claim of Froberg and Kane[23] that the TTO 'probably loses its advantage in large-scale studies due to its complexity and the resulting confusion and nonresponse'.

The substantive findings of this paper are that TTO valuations appear to be influenced by both the age and the sex of the respondent. Other background variables, such as social class and education, were found to be insignificant, and others (such as marital status), whilst statistically significant, are unlikely to be meaningful in programme evaluation. The importance of age and sex contradicts the findings of other studies: in their review of the literature to 1988, Froberg and Kane[24] find 'little compelling evidence of population differences due to demographic characteristics'. However, most of the previous studies of health state preferences have contained small numbers of respondents, and, as Froberg and Kane readily admit[24] 'low statistical power may be obscuring differences'. If genuine differences do exist between population sub-groups then the issue of *whose values should count* in clinical and public policy decision-making becomes an important one. In certain situations, it may be appropriate to weight more heavily the preferences of those most directly affected by an intervention or policy.

Before definite conclusions can be reached on this issue 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. Therefore, some of the differences in health state valuations reported in this paper may be the result of different perceptions of time rather than differences perceptions of severity of illness *per se*.

Despite the many unanswered questions, there is increasing recognition of the need to consider quality of life when making decisions at all levels. The results of this study show that eliciting the preferences of the general public is feasible and indicate possible directions for future research.

ACKNOWLEDGEMENTS

This Study was financed by the Department of Health. The survey work was conducted jointly with Social and Community Planning and Research. Bob Erens, Roger Thomas and Katarina Thomson were particularly influential in the study design, and we owe a tremendous debt to the field staff and their managers, and particularly to the small dedicated band of interviewers who co-operated so productively in the developmental work. The paper benefited greatly from the comments of two anonymous referees.

REFERENCES

1. Torrance, G. W. Measurement of health state utilities for economic appraisal, *Journal of Health Economics* 1986; **5**: 1–30.
2. Brooks, R. G. *Health status and quality of life measurement: issues and developments*, Swedish Institute of Health Economics, IHE, Lund, 1991.
3. Buckingham, K. and Drummond, M. *A theoretical and empirical classification of health valuation techniques*, HESG Conference, Strathclyde, 1993.
4. Richardson, J. Cost-utility analysis: what should be measured?, *Social Science and Medicine* 1994; **39**(1): 7–21.
5. Dolan, P., Gudex, C., Kind, P. and Williams, A., *Valuing health states: a comparison of methods*, HESG Conference, Strathclyde, 1993.

6. O'Brien, B. J. and Drummond, M. F. Statistical versus quantitative significance in the socioeconomic evaluation of medicines, *PharmacoEconomics* 1994; 5(5): 389–398.
7. Gudex, C. *Time trade-off user manual: Props and self-completion method*, Centre for Health Economics Occasional Paper Series, 1994.
8. Eyman, R. K. The effect of sophistication on ratio and discriminative scales, *American Journal of Psychology* 1967; **80**: 520–540.
9. Poulton, E. C. *Bias in quantifying judgements*, Lawrence Erlbaum, Hove, 1989.
10. Patrick, D. L., Starks, H. E., Cain, K. C., Uhlmann, R. F. and Pearlman, R. A. Measuring preferences for health states worse than death, *Medical Decision Making* 1994; **14**: 9–18.
11. Rosser, R. and Kind, P. A scale of valuations of states of illness: is there a social consensus? *International Journal of Epidemiology* 1978; **7**: 347–358.
12. Read, J. L., Quinn, R. J., Berwick, D. M., Fineberg, H. V. and Weinstein, M. C. Preferences for health outcomes: comparison of assessment methods, *Medical Decision Making* 1984; **4**: 315–329.
13. Sutherland, H. J., Llewellyn–Thomas, H., Boyd, N. F. and Till, J. E. Attitude toward quality of survival: the concept of maximal endurable time, *Medical Decision Making* 1982; **2**: 299–309.
14. Sackett, D. L. and Torrance, G. W. The utility of different health states as perceived by the general public, *Journal of Chronic Diseases* 1978; **31**: 697– 704.
15. Shiell, A., King, M. and Briggs, A. *The consistency of rating scale and time trade-off techniques for eliciting preference weights for health states*, HESG Conference, Strathclyde, 1993.
16. Kahneman, D., Knetsch, J.L. and Thaler, R. Experimental tests of the endowment effect and the Coase theorem, *Journal of Political Economy* 1990; **98**: 1325–48.
17. Dolan, P., Jones–Lee, M. and Loomes, G. Risk trade-off vs standard gamble procedures for measuring health state utilities, *Applied Economics* 1995; **27**: 1103–1111.
18. Boyd, N. F., Sutherland, H. J., Ciampi, A., Tibshirani, R., Till, J. E. and Harwood, A. A comparison of methods of assessing voice quality in laryngeal cancer, in *Choices in health care: decision-making and evaluation of effectiveness*, Department of Health Administration, University of Toronto, 1982.
19. Wolfson, A. D., Sinclair, A. J., Bombardier, C. and McGeer, A. Preference measurements for functional status in stroke patients: interrater and intertechnique comparisons, in Kane, R. L., Kane, R. A. (eds) *Values and long term care*, Lexicon Books, Mass, 1982.
20. Churchill, D. N., Torrance, G. W., Taylor, D. W., Barnes, C. C., Ludwin, D., Shimizu,A. and Smith, E. K. M. Measurement of quality-of-life in end-stage renal disease: the time trade-off approach, *Clinical and Investigative Medicine* 1987; **10**: 439–471.
21. Christensen–Szalanski, J. J. Discount functions and the measurement of patients' values: women's decisions during childbirth, *Medical Decision Making* 1984; **4**: 45–58.
22. Llewellyn–Thomas, H. A., Sutherland, H. J. and Thiel, E. C. Do patients' evaluations of a future health state change when they actually enter that state?, *Medical Care* 1993; **31**: 1002–1012.
23. Froberg, D. G. and Kane, R. L. Methodology for measuring health state preferences IV: progress and a research agenda, *Journal of Clinical Epidemiology* 1989; **42**: 675–685.
24. Froberg, D. G. and Kane, R. L. Methodology for measuring health state preferences III: population and context effects, *Journal of Clinical Epidemiology* 1989; **42**: 585–592.