University of Southern Denmark · Syddansk Universitet

An Assessment of Inconsistencies in the Valuation of Hypothetical EuroQol (EQ-5D) Health States

Kim U. Wittrup-Jensen[1,4], *Jørgen Lauridsen*[2], *Kjeld M. Pedersen*[3]

(1): Bayer HealthCare AG, kim.wittrup-jensen@bayerhealthcare.com

(2): Institute of Public Health – Health Economics, E-mail jtl@sam.sdu.dk

(3): Institute of Public Health – Health Economics, University of Southern Denmark, kmp@sam.sdu.dk

(4): The study was done when Kim U. Wittrup-Jensen was a PhD student at University of Southern Denmark

Health EconomicsPapers
2008:5

University of Southern Denmark

Abstract

Background: Valuation methods for measurement of Health-Related Quality of Life (HRQoL) are increasingly being used in medical decision-making contexts, e.g. by estimating Quality Adjusted Life-Years (QALYs). QALYs are estimated by using scaling methods such as the Visual Analogue Scale (VAS), Time Trade-Off (TTO) or the Standard Gamble (SG). However, all techniques, to some degree, suffer to some extent from the problem of inconsistent responses.

Objectives: To examine the degree to which three different preference-elicitation methods – ranking of health states, VAS, TTO – produces inconsistent responses, and the effect of socio-demographic and other health characteristics on the numbers of inconsistencies produced.

Data and methods: 4,074 addresses in Denmark, stratified by age, gender and geographic address, were contacted by phone. 1,332 interviews were conducted in the respondent's own ho-mes, where respondents were asked to assess hypothetical health states, generated within the EuroQol (EQ-5D) classification system, within three different exercises: (1) a ranking exercise, (2) VAS, and (3) TTO. In total, 46 health states were directly valued by respondents. A split-sample technique was applied where each respondent valued 14 to 16 different EQ-5D health states. One type of inconsistency and one type of consistency were tested. Internal (logical) inconsistency and criterion consistency were assessed by both the numbers of inconsistencies produced by each of the three valuation methods, and the 'size' of those inconsistencies, i.e. the distance in terms of severity between the given health states rated inconsistently was tested. Finally, by estimating Kendall's correlation coefficients it was possible to assess the impact of socio-demographic characteristics on the number, rate, and index of inconsistencies exhibited within all three exercises.

Results: The study shows that there is a high degree of criterion consistency between all three exercises; however, the highest is between the VAS and the TTO exercises. *Age* is a significant factor for criterion consistency as the degree of consistency decreases with age. Also *years spent in (primary) school* has a significant impact, where the degree of consistency increases with years spent in (primary) school. The number of internal (logical) inconsistencies increases across exercises; the lowest is in the ranking exercise and the highest in the TTO exercise. The number of *years spent in (primary) school* has a decreasing effect on both the rate and index of internal (logical) inconsistencies. This tendency is present across all three exercises.

Conclusions: Not surprisingly, due to its complexity compared to the ranking and the VAS, it is the TTO exercise where respondents exhibit the most internal (logical) inconsistencies. The results indicate that efforts should be made to reduce inconsistencies, particularly among elderly and lesser-educated respondents.

Introduction

An important issue within the measurement of health status involves the valuation of health states by means of different scaling techniques, e.g. Category Rating, Magnitude Estimation, Equivalence of Numbers, Graphical Rating Scales, Visual Analogue Scale (VAS), Time Trade-Off (TTO), and the Standard Gamble (SG) [Torrance 1986; Froberg & Kane 1989a]. Often, different people have different attitudes and preferences to questions concerning their daily living. Hence, it has to be expected that preferences for health states will vary across individuals. It is therefore not surprising that, within each scaling technique, certain significant differences may appear between different groups of the population [Froberg & Kane 1989b]. On the other hand there exist cases where the valuations have varied within different scaling techniques within the same group. There exist a number of documented cases where the variation in the answers is caused by the framing of the health states, i.e. the framing of the questions concerning the trade-off between health states, or by which methods have been applied [Torrance 1976; Nord 1992]. Studies have also documented that socio-demographic factors and health characteristics are important (and significant) factors in how individuals assess health states [Badia *et al.* 1998; Gudex *et al.* 1997; Dolan *et al.* 1996; Dolan 1996].

It is documented that the valuations of health states, when described in a natural and informal scenario, are lower than when described in a formal and standardized form [Llewellyn-Thomas et al. 1984]. In another survey it is shown that the SG method causes internal inconsistency, i.e. when the alternative is kept at a constant level changes in the outcome in the risk-emphasized alternative will influence the estimated values [Llewellyn-Thomas *et al.* 1982]. In both cases inconsistencies arise as a direct consequence of the method used as measurement. Within the literature, Kahnemann & Tversky refer to this form of inconsistency as *framing effects* [Kahnemann & Tversky 1981].

Inconsistencies that happen as a result of framing effects have to be accepted to a certain degree as most respondents, be they the general population, patients, or health care professionals, are rarely confronted with questions concerning their preferences for health, or for that matter, health care services, in the form illustrated in the above examples.

An example could be answering how may patients who suffer from a chronic disease have to be restored to perfect health in order to make this equivalent to restoring 10 patients, who otherwise would have died, to full health. The converse question could be formulated as the number of deaths one would accept in order to restore 10 patients, who suffers from a chronic disease, to full health. Insofar as preferences are assumed not to be dependent on point of reference, nothing has changed except by how the question is framed. According to *Prospect Theory* (PT) it is explicitly assumed that preferences are dependent on the point of reference: the result may differ according to how the question is framed, i.e. whether respondents are asked, on the one hand, about patients who will be rescued from dying and, on the other hand, about patients who are expected to die. In empirical studies it has been documented that PT is, to a certain degree, robust, i.e. the answers vary according to how the questions are framed [Prades 1997].

Inconsistencies that result from the respondent being unable to comprehend the exercise through inability to understand or mis-interpretation are more serious than inconsistencies caused purely by framing effects. These types of inconsistencies will, in situations where the respondent has been asked to value a given set of health states, show themselves as a ranking of preferences which does not correspond with *a priori* expectations. It is thus important to distinguish between *primary inconsistencies*, caused by the respondent's misunderstanding of the questions, and *secondary inconsistencies*, caused by methodological aspects in the measurement procedure [Dolan & Kind 1996].

The number of inconsistencies caused by primary factors is largely unknown. Potential explanations are often lost in the phase of processing empirical data, where this process is initially un-dertaken *after* the exclusion of the respondents who did not display the standard ranking put forward by the researcher. In a study by Bush *et al.* (1977) respondents who valued health states inconsistently with regard to the prior instructions were systematically excluded, which caused a loss of over twenty per cent of the data material. In another study by Sackett & Torrance (1978) respondents were only included if they had valued all health states according in an internally consistent way. The result was an exclusion of 22 per cent of the respondents in the data material. The application of the above or similar criteria to exclude respondents may have important implications for the choice of the HRQoL questionnaire, scaling method, administration of the study and, not least, the valuation of health states. Only by carefully studying those respondents who violate *a priori* expectations, and by extensively analysing why these primary inconsistencies occur, will one be able to contribute to a more thorough understanding of how to measure health.

Objectives

The objective is to examine the degree to which three preference-elicitation methods (ranking of health states, valuation of health states using the VAS and TTO techniques) produce inconsistent responses and the effect of socio-demographic and other health characteristics on the numbers of inconsistencies produced.

The data

Study subjects were 1332 respondents who participated in a study which had the main purpose of modelling EQ-5D tariffs using the TTO technique. The distribution, by socio-demographic characteristics among the respondents, is illustrated in Table 1. 58 per cent of the respondents were female and around 29 per cent of the respondents were 60 years or above. Over 40 per cent of the respondents lived together with one person in their households. Exactly 30 per cent had a higher primary education (matching a high school degree) and around 21 per cent spent no more than seven years in primary school. Regarding further education, around 40 per cent had some sort of academic education, of which around 14 per cent held a university degree. Almost half of the respondents had a monthly income before tax (in year 2000) below 14,999 DKK. Around 45 per cent had an income in the range of 15,000 – 29,999 DKK, leaving the remaining few per cent with an income somewhere above 30,000 DKK.

Table 1. Socio-demographic and health state characteristics of the study population: *n* (%), total sample (*n* = 1,332).

Socio-demographic data:	
Gender:	
Male	560 (42.0)
Female	772 (58.0)
Age:	
18 – 29	210 (15.8)
30 – 59	734 (55.1)
≥ 60	388 (29.1)
Persons in the household:	
1	280 (21.0)
2	544 (40.8)
3	218 (16.4)
4	210 (15.8)
5	67 (5.0)
≥ 6	13 (1.0)
Years in school (primary):	
7	281 (21.1)
8 – 9	225 (16.9)
10	427 (32.1)
Higher education	399 (30.0)
Education:	
Unskilled	45 (3.4)
Basic Vocational Education (1 year)	38 (2.9)
Apprenticeship	149 (11.2)
Other training of apprentices	233 (17.5)
Short advanced studies (1 year)	188 (8.9)
Intermediate advanced studies (< 3 years)	232 (17.4)
Long advanced studies (3-4 years)	119 (8.9)
Long advanced studies (> 4 years)	69 (5.2)
Other	329 (24.7)
Monthly income (before tax)[ab]*:*	
< 14,999 DKK	585 (47.9)
15,000 – 21,999 DKK	343 (28.1)
22,000 – 29,999 DKK	190 (15.6)
30,000 – 44,999 DKK	81 (6.6)
45,000 – 59,999 DKK	11 (0.9)
≥ 60,000 DKK	11 (0.9)

[a] 1 DKK = 0.1347 Euro.
[b] n = 1,221 due to missing observations.

In Table 2 the distribution on the EQ-5D descriptive system is illustrated, including the corresponding VAS scores illustrated by mean and median values. The majority of the respondents, within all five dimensions, reported that they had 'no problems'. However, over 30 per cent reported problems due to 'pain/discomfort'. The mean VAS score was 88.1 with a standard deviation of 15.2. Over 70 per cent of the respondents rated their own health status as being ≥ 90 and around 22 per cent rated themselves as having a score of 100 (i.e. perfect health).

Table 2. Distribution on EQ-5D descriptive system including VAS scores: *n* (%), total sample (*n* = 1,332).

EQ-5D descriptive system:	
Mobility	
Without problems	1,133 (85.1)
With problems	199 (14.9)
Self-care	
Without problems	1,292 (97.0)
With problems	40 (3.0)
Usual activities	
Without problems	1,115 (83.7)
With problems	217 (16.3)
Pain/discomfort	
Without problems	897 (67.3)
With problems	435 (32.7)
Anxiety/depression	
Without problems	1,193 (89.6)
With problems	139 (10.4)
VAS:	
Mean (SD)	88.1 (15.2)
Median	95.0
Maximum	100.0
Minimum	0.0
% of respondents with a score of ≥ 90	70.3
% of respondents with a score of 100	21.8

Note: "Without problems" applies to a score on the first of the three levels within dimension and "with problems" applies to a score either on the second or the third level.

Statistics

Due to the extensive programming required in order to calculate the different results of internal (logical) inconsistency criterion consistency, both the statistical programmes *SAS* and *STATA* were applied [SAS Institute Staff 2000; Rabe-Hesketh *et al.* 2000].

Methods

Inconsistency and consistency in the ranking, valuation, and TTO exercises

In order to assess the degree of (secondary) inconsistencies (and consistencies) one type of consistency and one type of inconsistency: c*riterion consistency* and *internal (logical) inconsistency* were defined. It was assumed that criterion consistency was present when: 1) the health states in the TTO exercise were ranked in the same way as they were in the ranking exercise, 2) the health states in the valuation exercise were ranked in the same way as they were in the ranking exercise, and 3) the health states in the TTO exercise were ranked in the same way as they were in the valuation exercise. A complete overview of the logic underlying the concept is illustrated in appendix A. The appendix contains four figures since the study was conducted as a split-sample design, where each sample was confronted with different EQ-5D health states.

Since the ranking exercise was the simplest of the three exercises, and because this exercise allowed the respondents to assess all health states at the same time, it was explicitly assumed that the ranking exercise was to be seen as the *gold standard* by which respondents assessed the health states within the study. In the assessment of criterion consistency between the valuation and TTO exercises, the ranking in the valuation exercise was assumed to be the *gold standard*.

Internal (logical) inconsistency was relevant within the ranking, valuation, and TTO exercises and was assessed in three forms: i) the *presence* or *absence* of inconsistency; ii) the *inconsistency rate* and, iii) as an *inconsistency index* [Badia *et al.* 1999].

Criterion consistency and internal (logical) inconsistency were investigated at both *individual* and *aggregated* levels.[1] In order to analyse internal (logical) inconsistency at the individual level the percentage of respondents, including confidence intervals with inconsistencies, mean (SD), and median for the rate of inconsistencies as well as for indicators of inconsistencies for the ranking, valuation, and TTO exercises were estimated. Differences in the inconsistency indicators were compared across respondents at the aggregated level, based on socio-demographic health characteristics including age, gender, number of people in the household, level of education and income.

At the individual level criterion inconsistency was assessed by the *Kendall coefficient*, between the ranking exercise and the TTO exercise, the ranking exercise and the valuation exercise, and the valuation exercise and the TTO exercise, respectively. The Kendall coefficient measures the degree of correlation between the mean rankings for the three exercises for all 46 health states. At the aggregated level the so-called 'robust methods' were used, applying a *median-regression method*[2], where the Kendall coefficients were the dependent variables and the socio-demographic respondent characteristics the explanatory variables.

Internal inconsistencies

The presence of inconsistencies was defined as the percentage of respondents who made inconsistencies. The inconsistency rate was the number of pair-wise inconsistently rated set of EQ-5D health states (expressed as a percentage) compared to all possible logical inconsistencies within all EQ-5D health states.[3] The inconsistency index was calculated based on the degree of severity, measured as the distance between the sets of health states which were ranked logically inconsistent. The index combined information on both the number of inconsistent answers and the distance between inconsistently valued health states. To calculate the (cardinal) distance between health states, we applied the Danish EQ-5D TTO-based set of tariffs [Wittrup-Jensen *et al.* 2001].

In order to illustrate how the inconsistency rate and the inconsistency index were calculated, one can imagine a given scaling instrument resulting in the following ranking of EQ-5D health states: 12211, 22123, 11122, 21223, 22223. In this ranking there is one logical inconsistency as the health state 22123 can never be better than the health state 11122. The total potential number of possible inconsistencies is six, i.e. 22223 better than 12211, 22223 better than 22123, 22223 better than 11122, 22223 better than 21223, 21223 better than 11122 and 22123 better than 11122. Let us assume that the respondent only displayed this one logical inconsistency out of the six possible inconsistencies, which results in an inconsistency rate of 0.167 (1/6).

What if it is assumed that it is more inconsistent to prefer health state 22223 than health state 11122, compared to health state 22123? This effect is captured in the inconsistency index. In order to estimate this index it is necessary to have an expression for the degree of severity, measured as the distance between the health states. In this particular example the following five arbitrarily chosen EQ-5D health states are given their corresponding Danish TTO-based tariffs, so that 12211 = 0.776, 22123 = 0.340, 11122 = 0.756, 21223 = 0.355, 22223 = 0.292. Hence, the total distance of possible inconsistencies is: (0.776 – 0.292) + (0.340 – 0.292) + (0.756 – 0.292) + (0.355 – 0.292) + (0.756 – 0.355) + (0.756 – 0.340) = 1.876. IF the distance between the health states which have been inconsistently ranked is 0.416 (0.756 – 0.340) the inconsistency index will be 0.222 (0.416/1.876). Both the inconsistency rate

[1] Aggregated level does not refer here to aggregated data, since all data are collected at the individual level, but merely that the respondents are viewed at the aggregated level according to their socio-demographic characteristics.

[2] The median-regression method is an estimation procedure that seeks to minimize the absolute residuals, rather than the sum of squares of the residuals as in ordinary least square regression. Median regression is less sensitive to outliers and skewness than is OLS. Since some respondents displayed no logical inconsistencies, this implies that the rate and index variables in Tables 7 & 8 are censored. This is an instance of quantile regression in which the 0.5 quantile (median) is used. It would be more appropriate to use a censored median-regression method (the so-called Powells CLAD estimation described in Johnston & Dinardo (1997)), instead of the uncensored as we do. However, this option is not standard within the *SAS* programme.

[3] In the literature the term *the city block distance* is also frequently used. [Fukada *et al.* 1999].

and the inconsistency index can fall within the interval 0 (no inconsistencies) and 1 (total inconsistency), which in this case would have to be multiplied by 100.

Results

Criterion consistency

In order to assess the degree of criterion consistency within the TTO exercise, it is necessary to use the ranking exercise in the TTO study as a basis and consider it as the *gold standard*. In Table 3, the degree of criterion consistency is illustrated. The Kendall correlation coefficients can take on values between ±1 and 0, where the correlation is higher the closer the coefficient is to ±1, and lower the closer to 0.[4] As can be seen, the correlation between the ranking exercise and the valuation exercise is relatively high, while it is a little lower between the ranking exercise and the TTO exercise and the valuation exercise and the TTO exercise.

Table 3. Criterion consistency, measured by Kendall correlation coefficients at the individual level.

	Mean (SD)	**Median**	**95 % CI**	***n***
Ranking versus valuation exercise	- 0.83 (0.23)	- 0.91	[- 0.84; - 0.82]	1,284
Ranking versus TTO exercise	- 0.60 (0.23)	- 0.65	[- 0.61; - 0.59]	1,276
Valuation versus TTO exercise	0.60 (0.23)	0.66	[0.59; 0.61]	1,323

Table 4 illustrates the degree of criterion consistency measured by socio-demographic indicators. The reasoning is that a negative variable is to be interpreted as a growing consistency. This means that the consistency between the ranking exercise and the valuation exercise increases with the number of years the respondents had 'spent in school', the 'number of people in the household' and 'income (before tax)'. On the other hand the consistency decreased with 'age' (however, this was very weakly significant). 'Gender' and 'education' had no effect. The consistency between the ranking exercise and the TTO exercise increased, likewise, with the 'number of people in the household' (very weakly significant), and with the number of 'years spent in school'. 'Income (before tax)' had a very small effect. The consistency was significantly lower for females and decreased with age. Furthermore, 'education' had no effect. In the last scenario the sign has to be interpreted in reverse as data were scaled equally within the valuation and TTO exercises. Hence the consistency between the valuation exercise and the TTO

[4] The sign is determined by the way the exercises are scaled. The Kendall correlation coefficients are negative between the ranking exercise and the TTO exercise, as the worse the health state is valued in the ranking exercise the higher the position, while the worse the health state is valued in the TTO exercise, the lower the position. The Kendall correlation coefficients are positive between the valuation exercise and the TTO exercise, since for both exercises the rule is: the worse the health state, the lower the position (value).

exercise increased with numbers of 'years spent in school', 'income (before tax)' and the 'number of people in the household' (very weakly significant). Females had a significantly lower consistency and the consistency tended to decrease with age. 'Education' had no effect.

Table 4. Criterion consistency explained by socio-demographic person characteristics; all coefficients are estimated based on all health states. Dependent variable = Kendall's Tau. Total sample (*n* = 1,332).

	Coefficient (S.E.)	*p*-value	95% confidence interval
Ranking → Valuation:			
Gender	-0.0023 (0.0089)	0.798	[-0.0198; 0.0152]
Age	0.0004 (0.0003)	0.131	[-0.0002; 0.0010]
Persons in the household	-0.0094 (0.0040)	0.019*	[-0.0172; -0.0016]
Years in school (primary)	-0.0037 (0.0021)	0.085**	[-0.0079; 0.0005]
Education	-0.0003 (0.0035)	0.932	[-0.0071; 0.0065]
Income (before tax)	-0.0008 (0.0005)	0.083**	[-0.0017; 0.0001]
Ranking → TTO:			
Gender	0.0234 (0.0117)	0.044*	[0.0006; 0.0463]
Age	0.0014 (0.0004)	0.000*	[0.0006; 0.0021]
Persons in the household	-0.0081 (0.0052)	0.118	[-0.0183; 0.0021]
Years in school (primary)	-0.0111 (0.0028)	0.000*	[-0.0166; -0.0056]
Education	0.0024 (0.0045)	0.597	[-0.0064; 0.0111]
Income (before tax)	-0,0009 (0.0006)	0.151	[-0.0020; 0.0003]
Valuation → TTO:			
Gender	-0.0194 (0.0106)	0.067**	[-0.0402; 0.0014]
Age	-0.0015 (0.0004)	0.000*	[-0.0021; -0.0008]
Persons in the household	0.0061 (0.0047)	0.196	[-0.0032; 0.0154]
Years in school (primary)	0.0124 (0.0026)	0.000*	[0.0073; 0.0174]
Education	-0.0025 (0.0040)	0.530	[-0.0104; 0.0054]
Income (before tax)	0.0013 (0.0005)	0.019*	[0.0002; 0.0023]

Note: *significant at p < 0.05, **significant at p < 0.10.

The level of internal (logical) inconsistency is assessed both in its *weak* and *strong* forms, where the weak form allowed health states to be valued at the same value and the strong form did not. From Table 5 it is seen that if the weak form is applied as a precondition for the number of inconsistencies the maximum number of inconsistencies displayed by any respondent was 26.[5] Over 20 per cent of the respondents did not display any inconsistencies and around 80 per cent had five or less inconsistencies. However, if the strong form is applied as a precondition, the picture changes. Now it is no longer 'legal' that two health states can have the same value and consequently the number of logical inconsistencies is higher. Only 2.2 per cent of the respondents displayed no inconsistencies, while around 27 per cent had five or less logical inconsistencies. Almost 40 per cent of the respondents had over 10 logical inconsistencies. The largest number of inconsistencies was 57, displayed by five respondents.

[5] The total number of possible (strong) inconsistencies which one respondent could display was 57.

Table 5. Number of strong and weak inconsistencies in the TTO exercise: *n* (%), total sample (*n* = 1,332).

Weak inconsistencies			Strong inconsistencies		
Pairwise inconsistencies	Number of responses (%)	Cumulative sum of responses (%)	Pairwise inconsistencies	Number of responses (%)	Cumulative sum of responses (%)
0	280 (21.0)	280 (21.0)	0	29 (2.2)	29 (2.2)
1	222 (16.7)	502 (37.7)	1	59 (4.4)	88 (6.6)
2	190 (14.3)	692 (52.0)	2	73 (5.5)	161 (12.1)
3	153 (11.5)	845 (63.5)	3	80 (6.0)	241 (18.1)
4	130 (9.8)	975 (73.3)	4	115 (8.6)	356 (26.7)
5	84 (6.3)	1,059 (79.6)	5	85 (6.4)	441 (33.1)
6	67 (5.0)	1,126 (84.6)	6	96 (7.2)	537 (40.3)
7	48 (3.6)	1,174 (88.2)	7	104 (7.8)	641 (48.1)
8	39 (2.9)	1,213 (91.1)	8	58 (4.4)	699 (52.5)
9	20 (1.5)	1,233 (92.6)	9	75 (5.6)	774 (58.1)
10	18 (1.4)	1,251 (94.0)	10	63 (4.7)	837 (62.8)
11-20	69 (5.2)	1,320 (99.2)	11-20	317 (23.8)	1,154 (86.6)
21-26	12 (0.8)	1,332 (100.0)	21-30	122 (9.2)	1,276 (95.8)
			31-40	33 (2.5)	1,309 (98.3)
			41-50	16 (1.2)	1,325 (99.5)
			51-57	7 (0.5)	1,332 (100.0)

Internal inconsistency (weak form) for those respondents who undertook the ranking, valuation, and TTO exercises are illustrated in Table 6. The table shows the percentage of logical inconsistencies that appeared at the individual level for all three exercises, the presence of internal inconsistency, the rate of internal inconsistency, and the index of internal inconsistency. In general, the number of logical inconsistencies was relatively high for all three exercises. The number of inconsistencies increased from exercise to exercise; around 60 per cent of the respondents displayed at least one inconsistency within the ranking exercise, while this number increased to around 80 per cent in the TTO exercise. With regard to the rate of logical inconsistencies, the numbers are to be interpreted as follows: out of all possible inconsistencies that can be displayed, around 4 per cent were ranked inconsistently within the ranking exercise. This number increased to around 8 per cent in the TTO exercise. However, it has to be said that 8 per cent was still quite a small number. The only form of variation occurred in the index of internal inconsistencies, where the TTO exercise resulted in the highest value, 4.76 versus 3.02 and 2.93 for the ranking and valuation exercises, respectively. These results indicate that health states assessed inconsistently within the TTO exercise had a tendency to be assessed differently with regard to the level of severity when the same health states are ranked (or valued) inconsistently within the ranking exercise (or valuation exercise).

Table 6. Internal (logical) inconsistencies at the individual level for the ranking, valuation, and TTO exercises: total sample (n = 1,332).

Types of internal inconsistency:	**Ranking exercise**	**Valuation exercise**	**TTO exercise**
Percentage of respondents with logical inconsistencies (%) [CI 95%]	60.96	66.59	78.62
	[58.34; 63.58]	[64.06; 69.13]	[75.93; 80.38]
Rate of logical inconsistencies (%)			
Mean (SD)	4.28 (7.97)	4.34 (7.23)	7.33 (8.56)
95 % confidence interval	[3.85; 4.71]	[3.96; 4.73]	[6.87; 7.79]
Median	2.33	2.33	4.87
Index of logical inconsistencies(%)			
Mean (SD)	3.02 (7.65)	2.93 (6.95)	4.76 (8.56)
95 % confidence interval	[2.61; 3.43]	[2.26; 3.31]	[4.30; 5.23]
Median	1.03	1.03	1.74

The *rate* of logical inconsistencies at the aggregate level explained by socio-economic characteristics is illustrated in Table 7. In the ranking exercise, the variables 'persons in the household' and 'years in school (primary)' were significant at the 5 per cent level. The coefficients are to be interpreted as follows: the more people there were in the respondent's household, the fewer tendencies the respondents had to display a logical inconsistency. The more the number of years spent in school, the lower the tendency for logical inconsistencies. In the valuation exercise, the variable 'years in school (primary)' was likewise significant. The variable 'age' was also significant; the tendency to display logical inconsistencies increased with age. As in the ranking and valuation exercises, the variable 'persons in the household' was significant in the TTO exercise. Furthermore, the variables 'years in school (primary)' and 'income (before tax)' were significant: the higher the income the lower the tendency to display logical inconsistencies.

Table 7. Rate of internal (logical) inconsistencies explained by socio-demographic characteristics; all coefficients estimated based on all health states. Dependent variable = Kendall's Tau. Total sample (*n* = 1,332).

	Coefficient (S.E.)	***p*-value**	**95 % confidence interval**
Ranking exercise:			
Gender	-0.0014 (0.0009)	0.112	[-0.0031; 0.0003]
Age	0.0000 (0.0000)	0.353	[0.0000; 0.0001]
Persons in the household	-0.0007 (0.0004)	0.066**	[-0.0015; 0.0001]
Years in school (primary)	-0.0005 (0.0002)	0.011*	[-0.0009; -0.0001]
Education	0.0001 (0.0003)	0.710	[-0.0005; 0.0008]
Income (before tax)	0.0000 (0.0030)	0.975	[-0.0001; 0.0001]
Valuation exercise:			
Gender	-0.0002 (0.0004)	0.669	[-0.0008; 0.0005]
Age	0.0000 (0.0000)	0.071**	[0.0000; 0.0000]
Persons in the household	-0.0002 (0.0002)	0.336	[-0.0005; 0.0002]
Years in school (primary)	-0.0002 (0.0001)	0.032*	[-0.0004; 0.0000]
Education	0.0000 (0.0001)	0.869	[-0.0003; 0.0002]
Income (before tax)	0.0000 (0.0000)	0.805	[0.0000; 0.0000]
TTO exercise::			
Gender	0.0001 (0.0041)	0.982	[-0.0079; 0.0081]
Age	-0.0001 (0.0001)	0.634	[-0.0003; 0.0002]
Persons in the household	-0.0031 (0.0018)	0.092**	[-0.0066; 0.0005]
Years in school (primary)	-0.0022 (0.0010)	0.025*	[-0.0041; -0.0003]
Education	0.0023 (0.0016)	0.145	[-0.0008; 0.0053]
Income (before tax)	-0.0001 (0.0002)	0.004*	[-0.0010; -0.0002]

Note: *significant at $p < 0.05$, **significant at $p < 0.10$.

Table 8. Index of internal (logical) inconsistencies explained by socio-demographic characteristics; all coefficients estimated based on all health states. Dependent variable = Kendall's Tau. Total sample (n = 1,332).

	Coefficient (S.E.)	*p*-value	95 % confidence interval
Ranking exercise:			
Gender	-0.0028 (0.0015)	0.060**	[-0.0057; 0.0001]
Age	0.0000 (0.0001)	0.987	[-0.0001; 0.0001]
Persons in the household	-0.0011 (0.0007)	0.089**	[-0.0024; 0.0002]
Years in school (primary)	-0.0011 (0.0004)	0.002*	[-0.0018; -0.0004]
Education	-0.0004 (0.0006)	0.476	[-0.0015; 0.0007]
Income (before tax)	0.0000 (0.0052)	0.688	[-0.0002; 0.0001]
Valuation exercise:			
Gender	-0.0004 (0.0016)	0.816	[-0.0036; 0.0028]
Age	0.0000 (0.0001)	0.660	[-0.0001; 0.0001]
Persons in the household	-0.0011 (0.0007)	0.144	[-0.0025; 0.0004]
Years in school (primary)	-0.0006 (0.0004)	0.120	[-0.0014; 0.0002]
Education	-0.0006 (0.0006)	0.348	[-0.0018; 0.0006]
Income (before tax)	0.0000 (0.0001)	0.811	[-0.0002; 0.0001]
TTO exercise:			
Gender	0.0004 (0.0023)	0.880	[-0.0042; 0.0049]
Age	-0.0001 (0.0000)	0.411	[-0.0002; 0.0001]
Persons in the household	-0.0009 (0.0010)	0.406	[-0.0029; 0.0012]
Years in school (primary)	-0.0020 (0.0006)	0.001*	[-0.0031; -0.0009]
Education	0.0001 (0.0009)	0.910	[-0.0016; 0.0018]
Income (before tax)	-0.0002 (0.0001)	0.102	[-0.0004; 0.0000]

Note: *significant at $p < 0.05$, **significant at $p < 0.10$.

Table 8 illustrates the *index* of logical inconsistencies explained by socio-demographic characteristics. In the ranking exercise, the variables 'gender', 'persons in the household' and 'years in school (primary)' were significant. As male was coded as 1 and female as 2, the coefficients show that males had a higher tendency for a high index rate than females. In other words, males had more 'serious' inconsistencies – measured on the degree of severity – than did females. The more persons in the household, the lower the inconsistency index and also the more years in school, the lower the index rate. There were no significant variables in the valuation exercise. In the TTO exercise only one variable was significant, namely 'years in school (primary)'.

Discussion

Looking at the results, it is a little surprising that there are so few respondents with zero (logical) inconsistencies and so relatively many respondents with many inconsistencies: nearly forty per cent displayed 10 inconsistencies or more. There is no straight forward explanation for this. As reported by Badia *et al.* (1999), the study also reports that the percentage of internal (logical) inconsistencies increased across exercises: the lowest within the ranking exercise and the highest within the TTO exercise. In total around 60 per cent of the respondents displayed at least one *weak* inconsistency in the ranking exercise, whereas nearly 80 per cent had at least one inconsistency in the TTO exercise. It was also found that the rate of logical inconsistencies increased across exercises, with a substantial (and significant at the 1 per cent level) increase from the va-luation exercise to the TTO exercise. Finally, the rate of internal (logical) inconsistencies drop-ped in the valuation exercise compared to the ranking exercise and then increased in the TTO exercise. The increase from both the ranking and valuation exercises to the TTO exercise was significant at the 1 per cent level. The percentage of logical inconsistencies and rate of logical inconsistencies reported by Badia *et al.* (1999) were much lower than our findings. On the other hand the index of logical inconsistencies in Badia *et al.* (1999) was a little different, since the index in the ranking exercise was higher than in our findings and the indexes in the valuation and TTO exercises were more or less similar. The reason for the differences between our findings and the findings reported by Badia *et al.* (1999) are manifold. The important thing, however, is that the number of logical inconsistencies increased as the exercise became more complex.

The study also report the level of criterion consistency by explicitly making the ranking exercise the *gold standard* exercise, the main reason being that the ranking exercise is the easiest and simplest of the three exercises, i.e. the exercise where the lowest number of logical inconsistencies was displayed. The findings show that there is a high degree of criterion consistency between the ranking and valuation exercises. This is not surprising since the respondents undertake the valuation exercise with the rank ordering from the ranking exercise still in front of them. The criterion consistency between the ranking exercise and the TTO exercise was still high at 0.60, however, this was lower than between the ranking and valuation exercises. Our findings point in the direction that while the ranking and valuation exercises are more or less the same, the TTO exercise is very different. The findings were similar to those reported by Badia *et al*. (1999).

It is important to investigate whether any socio-demographic characteristics (both assessed at the individual and the aggregated level) are indicators for the presence (or number) of inconsisten-cies. We investigated criterion consistency explained by socio-demographic characteristics. Again, we explicitly used the ranking exercise as a *gold standard*. The consistency between the ranking and valuation exercises increased with the number of years the respondents had spent in (primary) school, the number of individuals in the household, and income (before tax). Consistency decreased with age, however this was only very weakly significant (p = 0.131). Gender and education had no effect. The consistency between the ranking and TTO exercises increased likewise with the number of individuals in the household (p = 0.118) and the number of years spent in (primary) school. Income (before tax) had a very low effect (p = 0.151). The consistency was significantly lower for females and decreased with age. In addition, age had no effect. The consistency between the valuation and TTO exercises increased with the number of years spent in (primary) school, income (before tax), and the number of individuals in the household (p = 0.196). Females had a significantly lower consistency and the consistency tended to decrease with age. Education had no effect.

In summing up, is seems that criterion consistency decreases as individuals get older because they display more inconsistencies. In addition, criterion consistency appears to increase with the number of individuals and years spent in school. On the other hand education does not influence criterion consistency. Finally, criterion consistency seems to increase with income (before tax), pointing in the direction that poor individuals display more inconsistencies than rich individuals. Badia *et al.* (1999) also found that age was a significant factor in causing more inconsistencies at the individual level. However, their results cannot support our findings that the level of education is insignificant. This may be caused by differences in defining the level of education.

Finally, the study also looked at whether social characteristics at the aggregated level meant something within all three exercises. In the estimation of the rate of logical inconsistencies in the ranking exercise the variables 'persons in the household' and 'years in school (primary)' were significant at the 5 % level. The interpretation is that the more individuals there are in the respondent's household, the more the respondent has a tendency to display fewer logical inconsistencies. The higher the number of years spent in school, the lower the tendency to display logical inconsistencies. In the valuation exercise, the variable 'years in school (primary)' is likewise significant. Age is also significant – the tendency to display logical inconsistencies increases with age. As in the ranking and valuation exercises, the variable 'persons in the household' is also significant in the TTO exercise. In addition, the variables 'years in school (primary)' and 'income (before tax)' are significant – the higher the income, the lower the tendency to display logical inconsistencies. It seems that the rate of internal inconsistencies decreases with the number of individuals in the household and years spent in (primary) school. For the index of logical inconsistencies in the ranking exercise the variables 'gender', 'persons in the household' and 'years in school (primary)' are significant. Also males have a tendency to have a higher index than females. The interpretation is that males display more serious inconsistencies than females. Furthermore, the more individuals in the household the lower the inconsistency index, and also the more years in school, the lower the index rate. There are no significant variables in the valuation exercise. Finally, in the TTO exercise only one variable is significant – 'years in school (primary)'. To summarize, at the aggregated level, it appears that if there are more individuals in the household, the lower the tendency to have either a high rate or index within the ranking exercise. In the TTO exercises the number of years spent in (primary) school has a significant influence whether either the rate or index of internal inconsistencies is high.

In conclusion and based on the findings presented here, there is a need for reducing inconsistencies should be increased. Focus should be concentrated especially upon elderly respondents and those who have spent little time in (primary) school. Particularly for the TTO method respondents need to be allowed sufficient time to practise the (TTO) exercise. Perhaps also inconsistent answers could be pointed out to respondents and they could be allowed to revise such answers. Badia *et al.* (1999) reported that interviewers had a significant impact on the number of inconsistencies. We do not, however, have the data to address this issue, but nevertheless we would like to stress the importance of using experienced interviewers. According to Dolan (1996) current health status has an important effect on the valuations attached to different health states, with those in poorer health generally giving higher valuations. This finding poses problems for decision-makers since to the problem of whose values should count can be added the problem of when these values count, as the results (found by Dolan) imply that different valuations may be given by the same individual depending on how recent is his/her experience of illness. If it is believed that the word of the general population should count there is no problem since a randomised and representative sample will eliminate these problems. However, if it should be the word of the patients that is to count, we believe that Dolan (1996) has a point.

References

1. Badia X, Herdman M, Kind P. The influences of ill-health experience on the valuation of health. *Pharmaeconomics* 1998; 6: 687-96.
1. Badia X, Roset M, Herman M. Inconsistent responses in three preference-elicitation methods for health states. *Social Science and Medicine* 1999b; 49: 943-950.
2. Bush JW, Kaplan RM, Berry CC. Comparison of methods for measuring social preferences for a health state index. American Statistical Association Annual Meeting 1997: 682 -.
3. Dolan P. The effect of experience of illness on health state valuations. *Journal of Clinical Epidemiology* 1996; 49: 551-64.
4. Dolan P, Gudex C, Kind P, Williams A. Valuing health states: A comparison of methods. *Journal of Health Economics* 1996a; 15: 209-231.
5. Dolan P & Kind P. Inconsistency and health state valuations. *Social Science and Medicine* 1996; 4: 609-615.
6. Froberg DG & Kane RL. Methodology for measuring health-state preferences – II: Scaling methods. *Journal of Clinical Epidemiology* 1989a; 42: 459-471.
7. Froberg DG & Kane RL. Methodology for measuring health state preferences – III: Population and context effects. *Journal of Clinical Epidemiology* 1989b; 42: 585.
8. Fukada T, Hamashima C, Hisashige A, Ikeda S, Ikegami N, Nishimura S, Sakai I, Tamura M, Tsuchiya A. *Estimating a EuroQol value set: The case of Japan*. Discussion Paper presented at the EuroQol Group meeting in 1999, Sitges, Spain.
9. Gudex C, Dolan P, Kind P, Williams A. Health state valuations from the general public using the visual analogue scale. Quality of Life Research 1997; 5: 521-31.
10. Johnston J & DiNardo J. *Econometric Methods*. London: McGraw-Hill 1997.
11. Kahnemann D & Tversky A. The framing of decisions and the psychology of choice. *Science* 1981; 211: 453.
12. Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF. The measurement of patients' values in medicine. *Medical Care* 1982; 17: 501.
13. Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF. Describing health states: Methodologic issues in obtaining values for health states. *Medical Care* 1984; 22: 543.
14. Nord E. Methods for quality adjustment of life years. *Social Science and Medicine* 1992; 34: 559-69.
15. Prades JLP. Is the person trade-off a valid method for allocating health care resources? *Health Economics* 1997; 6: 71-81.
16. Rabe-Hesketh S & Everitt B. *A handbook of statistical analyses using STATA*. New York: CRC Press. 2000.

17. Sackett DL & Torrance GW. The utility of different health states as perceived by the general public. *Journal of Chronic Disorders* 1978; 31: 697 -705.
18. SAS Institute Staff. *SAS/STAT User's guide Version 8*. New York: Fatbrain. 2000.
19. Torrance GW. Social preferences for health states: An empirical evaluation of three measurement techniques. *Socio-Economic Planning Science* 1976; 10: 129-36.
20. Torrance GW. Measurement of health state utilities for economic appraisal. A review. *Journal of Health Economics* 1986; 5: 1-30.

Appendix A

Figure A.1. Logical consistencies in split-sample I.

i \ j	23321	11211	11212	21232	12211	11113	32331	22333	22233	22222	21111	21222	33333	23332
23321	-	≤	?	?	≤	?	?	?	?	?	≤	?	≥	?
11211		-	≥	≥	≥	?	≥	≥	≥	≥	?	≥	≥	≥
11212			-	≥	?	?	?	≥	≥	≥	?	≥	≥	≥
21232				-	?	?	?	≥	≥	?	≤	≤	≥	≥
12211					-	?	≥	≥	≥	≥	?	?	≥	≥
11113						-	?	≥	≥	?	?	?	≥	?
32331							-	?	?	?	≤	?	≥	?
22333								-	≤	≤	≤	≤	≥	?
22233									-	≤	≤	≤	≥	?
22222										-	≤	≤	≥	≥
21111											-	≥	≥	≥
21222												-	≥	≥
33333													-	≤
23232														-

Figure A.2. Logical consistencies in split-sample II.

i \ *j*	32211	11112	12222	21312	33322	22323	12223	11131	12111	22222	22121	33333	21133	21221
32211	-	?	?	?	?	?	?	?	≤	?	?	≥	?	?
11112		-	≥	≥	≥	≥	≥	?	?	≥	?	≥	≥	?
12222			-	?	≥	≥	≥	?	≤	≥	?	≥	?	?
21312				-	≥	≥	?	?	?	?	?	≥	?	?
33322					-	?	?	?	≤	≤	≤	≥	?	≤
22323						-	≤	?	≤	≤	≤	≥	?	≤
12223							-	?	≤	?	?	≥	?	?
11131								-	?	?	?	≥	≥	?
12111									-	≥	≥	≥	?	?
22222										-	≤	≥	?	≤
22121											-	≥	?	?
33333												-	≤	≤
21133													-	?
21221														-

Figure A.3. Logical consistencies in split-sample III.

j / *i*	21121	11211	21322	32223	22331	13311	32313	22222	22112	11121	22122	23313	33333	11122
21121	-	?	≥	≥	≥	?	?	≥	?	≤	≥	?	≥	?
11211		-	≥	≥	≥	≥	≥	≥	?	?	?	?	≥	?
21322			-	?	?	?	?	?	?	≤	?	?	≥	≤
32223				-	?	?	?	≤	≤	≤	≤	?	≥	≤
22331					-	?	?	?	?	≤	?	?	≥	?
13311						-	?	?	?	?	?	≤	≥	?
32313							-	?	≤	?	?	?	≥	?
22222								-	≤	≤	≤	?	≥	≤
22112									-	?	≥	≥	≥	?
11121										-	≥	?	≥	≥
22122											-	?	≥	≤
23313												-	≥	?
33333													-	≤
11122														-

Figure A.4. Logical consistencies in split-sample IV.

i \ *j*	11112	22322	11221	33232	32232	11133	33321	21323	13212	22222	BEV[1]	21111	33333	11312
11112	-	≥	?	≥	≥	≥	?	≥	≥	≥	?	?	≥	≥
22322		-	≤	≥	?	?	?	?	?	≤	?	≤	≥	≤
11221			-	≥	≥	?	≥	≥	?	≥	?	?	≥	?
33232				-	≤	?	?	?	≤	≤	?	≤	≥	?
32232					-	?	?	?	?	≤	?	≤	≥	?
11133						-	?	?	?	?	?	?	≥	?
33321							-	?	?	?	?	≤	≥	?
21323								-	?	?	?	≤	≥	≤
13212									-	?	?	?	≥	?
22222										-	?	≤	≥	?
UNC[1]											-	?	?	?
21111												-	≥	≥
33333													-	≤
11312														-

Note: [1] Unconscious.