# INCONSISTENCIES AND MODELLING OF THE FINNISH EUROQOL (EQ-5D) PREFERENCE VALUES

by Ohinmaa Arto and Sintonen Harri

## 1 INTRODUCTION

Most of the EQ-5D states can be ordered logically in terms of relative goodness (or badness) of the states compared to other states. When people are assigning values to different health states, they may violate this logical order, which can be seen as inconsistencies. The health state which is 'at least as good' on all EQ-5D dimensions should logically get a score at least as high as the health state to which it is compared (Dolan and Kind 1996, Johnson et al. 1998). Inconsistencies may arise for various reasons. For example, the respondent may not have read the health state descriptions carefully enough to recognise a logical ordering between them or he/she may not have understood the valuation task right.

The existence of inconsistencies may potentially constitute a serious problem when inferences are made on the basis of individual valuations on the relative value of different health states at a group/societal level. In the EuroQol context a number of health states are valued directly by a sample of individuals and these data are due to be used to derive a societal value for these states and to all other EuroQol states not directly valued, i.e. in deriving a so called tariff or set of single index scores for the health states. The purpose of the individual valuation data are thus to provide reliable and valid raw material for modelling a tariff.

The purpose of this paper is to explore the existence and frequency of inconsistencies in the Finnish valuation data and the extent to which they affect the modelling of the data (parameter estimates and fit) and the resulting tariff values. This is hoped to provide some indication on how to best deal with such inconsistencies when trying the elicit a valid tariff through the modelling approach.

# 2 METHODS

## 2.1 DATA

The Finnish EuroQol valuation questionnaires were mailed in November 1992 to a sample of 4000 chosen randomly from the national computerised population registry. The respondents were over 16 years of age and in the sample the elderly were overestimated due to expected poorer response rate and lower frequency in the population. The sample was divided into 17 sub-samples that received a different questionnaire. This study is based on 11 sub-samples (N=2530) utilising the standard EQ-5D VAS valuation method and covering altogether 43 EQ-5D states (plus the states of being dead and unconscious).

The EQ-5D measures health in 5 dimensions: mobility (mob), self-care (sc), usual activities (ua), pain or discomfort (pain), and anxious or depressed (mood). Each dimension has three levels of severity: no problems (coded as 1), some/moderate problems (coded as 2) and severe/extreme problems (coded as 3), so that each health state can be represented by a five digit number ranging from 11111 (best possible health state) to 33333 (worst possible health state). Some HRQOL states were common to each sub-sample questionnaire, e.g. states 11111, 33333 and being dead, while some states were unique to various sub-samples (see Table 2). Two reminders were mailed about two weeks apart.

## 2.2 EXCLUSIONS AND INCONSISTENCIES

The difficulty with completing the questionnaire was measured by a direct four-point question and by an estimate of completion time. Indirect evidence of the difficulty can also be obtained from the 'cleaning' of the data by using several criteria. The criteria for excluding observations from the data were as follows:
- the respondent had valued less than 3 health states, or
- all states were valued the same (±5), or
- state 11111 was not valued, or
- the state of being dead was valued higher or equal to state 11111.

The order of the endpoints of the scale (11111 and death) is important valuewise, since the value of the lower end of the scale cannot exceed that of the best EQ-5D state.

In the study, a state was considered inconsistent if it had a higher preference value than at least one higher rank (logically better) state (e.g. 11122 > 11112). The value of three states (11111, 33333, and dead) was given on two pages. In this parametric test, the higher value of states 11111, and the lower value of states 33333 and the dead was used in the comparison to

other states. For exploring the effects of inconsistencies the 'cleaned' final sample (n=1272) was divided into four inconsistency groups: respondents with no inconsistencies, one state

with at least 1 inconsistency, 2-3 states with at least one inconsistency, and from 4 to 12 states with inconsistencies.

## 2.3 MODELLING

The modelling was based on the Ordinary Least Squares (OLS) technique. The three levels of a dimension were coded using a separate dummy variable for levels 2 and 3 on each dimension. In addition separate dummy variables were used to indicate the difference from the best state (11111) to whatever state (constant), to the states of being dead and unconscious, and a N3 dummy was included to measure the possible shift when a state included at least one dimension at level 3 (Dolan 1997). Interaction terms between the dimensions were not estimated, because the earlier EQ-5D studies have not found them statistically significant (Dolan et al. 1996, Badia et al. 1998). The estimated model is as follows:

$$PV = 11111 - \beta_0*constant - \beta_1*mob2 - \beta_2*mob3 - \beta_4*sc2 - \beta_5*sc3 - \beta_5*ua2 - \beta_6*ua3 - \beta_7*pain2 - \beta_8*pain3 - \beta_9*mood2 - \beta_{10}*mood3 - \beta_{11}*unconscious - \beta_{12}*N3 - \beta_{13}*death - e, \quad (1)$$

where PV is preference value for a state, β's are parameter estimates, and e is the error term.

The models utilised both individual data and logistic transformations of individual data (Abdalla & Russell 1995, Dolan 1995). In this study two transformations were used:

(2) the logit function $g_1 = \ln (M /(1 - M))$,

(3) the complementary log-log function $g_2 = \ln ( -\ln (1 - M))$, where

M stand for a QOL state scaled from 0.05 to 0.95. The purpose of the transformations was to normalise the observed preference values. In the earlier Finnish study the log-log function (formula 3) proved to give nearly normally distributed values for the mean values of the health states in a sub-sample of this data set (Ohinmaa et al. 1996).

The OLS models estimated the parameter values for the anchor points of the scale and these values were used to transform the EQ-5D tariff values to 0 (being dead) and 1 (11111) scale. In the logistic models the estimated values were first transformed back to a 0-1 scale and then the anchor points were used to produce the single index weights.

VIF test was used to test the multicollinearity of the model parameters. The specification of the models was tested for nonlinearity by Reset test and for heteroscedasticity by Breusch-Pagan (BP) test (Kmenta 1992).

## 3 RESULTS

With two reminders 1634 respondents returned the questionnaire (64.5 %). The background characteristics of the respondents are shown by gender in Table 1.

The mean age of male respondents was 51 years and that of females 50 years with much the same distribution into the age groups (Table 1). The females were slightly better educated than males and they had more experience of serious illness in family and when caring others than males. About 38% of males regarded the completion of the questionnaire as fairly or very difficult in contrast to 50% of females (P=0.0001). However, the mean completion time and also the percentage of included respondents after applying the exclusion criteria were the same for both genders (Table 1).

The self-rated health of men was slightly worse than that of women on the VAS scale. There were no statistically significant differences between the genders in neither the EQ-5D dimensions or in a variable that showed whether a respondents had any health problems on any of the five dimensions or not (Table 1).

Altogether 362 (22%) respondents were excluded on the basis of various criteria (Table 1). The elderly were more likely to be excluded from the data. Of the respondents younger than 45 years about 93% fulfilled the inclusion criteria, while in the elderly (65-95 years) 62% were included in the data. Also the respondents with low basic education had a tendency to be excluded more often, although that effect was statistically significant only in the age groups between 35 and 54 years. The gender did not have any effect on the exclusion.

**Table 1. Basic background variables of the sample by gender as percentages or mean values and standard deviations (N=1634).**

| Variables | Male (n=796) | Female (n=838) | P-value |
|---|---|---|---|
| Age: 17-24 | 11.6 | 12.9 | ns. |
| 25-34 | 15.1 | 17.9 | |
| 35-44 | 14.8 | 13.8 | |
| 45-54 | 11.2 | 11.8 | |
| 55-64 | 9.7 | 9.5 | |
| 65-74 | 25.4 | 20.8 | |
| 75-95 | 12.2 | 13.2 | |
| Basic education: | | | |
| Elementary school | 54.5 | 46.4 | 0.0007 |
| Basic school | 28.8 | 30.1 | |
| High school | 16.7 | 23.5 | |
| Vocational education: | | | |
| None or short courses | 59.1 | 55.7 | ns. |
| 2 year schooling | 19.2 | 17.8 | |
| Polytechnic | 14.6 | 17.7 | |
| University | 7.1 | 8.8 | |
| Experience of serious illness: | | | |
| Self (% yes) | 35.9 | 33.4 | ns. |
| In family (% yes) | 42.9 | 53.9 | 0.0002 |
| When caring others (% yes) | 23.4 | 44.2 | 0.0000 |
| Completion of questionnaire: | | | |
| Very difficult | 7.9 | 10.5 | 0.0000 |
| Fairly difficult | 30.5 | 39.1 | |
| Fairly easy | 47.6 | 41.3 | |
| Very Easy | 14.0 | 9.1 | |
| Mean completion time, min. | 18.4 (sd. 17.8) | 19.2 (sd. 18.5) | ns. |
| Inclusion of responses: | | | |
| Included (n=1272) | 76.4 | 79.2 | ns. |
| Valued less than 3 states | 12.9 | 11.1 | |
| All states valued the same | 4.1 | 2.4 | |
| State 11111 not valued | 3.7 | 5.1 | |
| Value of dead ≥11111 | 2.9 | 2.2 | |
| Mean of own health on VAS | 75.8 (sd. 19.0) | 78.0 (sd. 18.8) | 0.0246 |
| At least 1 health problem (%) | 52.9 | 50.7 | ns. |

Of all respondents 616 (38%) did not value the state of being dead and if they were also excluded from the data, 946 (58%) respondents of the original 1634 could be included in the Finnish VAS valuation data. This study utilises mostly the data set of 1272 respondents, because the value for the state of being dead was not used to transform the individual data before statistical analysis or single index modelling.

**Table 2. Mean and median values of the health states together with a rate of respondents valuing the state inconsistently with at least one other health state (N=1634).**

| State | Mean | Median | S.D | N | Inconsistency rate |
|---|---|---|---|---|---|
| 11111A | 89.93 | 96 | 15.29 | 1384 | (0.260)* |
| 11111B | 90.18 | 97 | 15.86 | 1353 | (0.237)* |
| 11112 | 65.13 | 70 | 22.39 | 761 | 0.113 |
| 11121 | 75.50 | 80 | 18.29 | 756 | 0.114 |
| 11122 | 49.31 | 50 | 22.91 | 750 | 0.077 |
| 11211 | 78.94 | 80 | 18.31 | 767 | 0.116 |
| 11233 | 46.33 | 40 | 29.89 | 133 | 0.376 |
| 11312 | 57.14 | 55 | 24.96 | 251 | 0.251 |
| 12111 | 63.97 | 70 | 24.22 | 774 | 0.039 |
| 12112 | 63.26 | 65 | 23.49 | 232 | 0.397 |
| 12121 | 54.93 | 60 | 24.31 | 228 | 0.206 |
| 12122 | 57.95 | 58 | 23.02 | 264 | 0.507 |
| 12211 | 70.33 | 70 | 20.60 | 254 | 0.378 |
| 12222 | 40.84 | 40 | 22.90 | 249 | 0.177 |
| 12223 | 47.72 | 45 | 27.60 | 227 | 0.476 |
| 13233 | 39.00 | 35 | 25.70 | 130 | 0.277 |
| 13332 | 42.53 | 37 | 28.42 | 239 | 0.309 |
| 21111 | 77.4 | 80 | 17.67 | 783 | 0.114 |
| 21133 | 40.29 | 35 | 26.88 | 228 | 0.259 |
| 21222 | 39.38 | 40 | 22.08 | 263 | 0.163 |
| 21232 | 42.57 | 40 | 23.05 | 751 | 0.285 |
| 22111 | 71.17 | 72 | 19.09 | 251 | 0.191 |
| 22112 | 53.60 | 50 | 19.88 | 242 | 0.165 |
| 22121 | 57.25 | 60 | 19.76 | 247 | 0.206 |
| 22122 | 44.47 | 45 | 20.87 | 243 | 0.296 |
| 22211 | 56.21 | 60 | 21.02 | 234 | 0.201 |
| 22212 | 51.50 | 50 | 21.32 | 240 | 0.279 |
| 22221 | 46.14 | 49 | 22.92 | 232 | 0.362 |
| 22222 | 37.33 | 35 | 21.73 | 238 | 0.286 |
| 22233 | 36.42 | 30 | 25.31 | 733 | 0.445 |
| 22323 | 29.75 | 25 | 23.23 | 770 | 0.265 |
| 22331 | 43.08 | 40 | 24.55 | 251 | 0.275 |
| 23232 | 44.80 | 40 | 26.04 | 244 | 0.529 |
| 23321 | 39.40 | 39 | 22.88 | 241 | 0.092 |
| 31231 | 43.23 | 40 | 24.77 | 235 | 0.111 |
| 31323 | 40.01 | 35 | 27.74 | 236 | 0.25 |
| 32132 | 36.56 | 30 | 25.69 | 256 | 0.254 |
| 32211 | 47.40 | 45 | 24.02 | 769 | 0.269 |
| 32223 | 29.60 | 25 | 23.25 | 264 | 0.348 |
| 32232 | 28.87 | 20 | 24.27 | 239 | 0.297 |
| 32313 | 26.76 | 19 | 26.06 | 233 | 0.155 |
| 33122 | 26.00 | 20 | 25.63 | 240 | 0.108 |
| 33321 | 27.37 | 20 | 25.74 | 737 | 0.137 |
| 33323 | 18.24 | 10 | 23.17 | 259 | 0.208 |

| | | | | | |
|---|---|---|---|---|---|
| 33333A | 17.56 | 5 | 26.23 | 1317 | 0.226 |
| 33333B | 17.46 | 6 | 26.01 | 1306 | 0.228 |
| DeadA | 14.96 | 4 | 24.00 | 1040 | - |
| DeadB | 14.68 | 4 | 23.58 | 1039 | - |
| Unconscious | 13.72 | 1 | 26.65 | 717 | - |

* state 11111 valued lower than at least one other state

The mean values of all respondents were systematically lower than median values in the upper end of the scale while for the states valued closer to 0 the means tended to exceed the medians (Table 2). The maximum of all states were 100 and the minimum was 0 for all other states except state 11312 (value of 2). Table 2 presents also the state-specific inconsistency rates, which show the percentage of respondents valuing the health state inconsistently with at least one other health state. In this parametric test, the inconsistency rate differed considerably depending on the seriousness of the health state. Very mild states had less inconsistencies than more serious ones. Although the state 33333 has the lowest logical rank and thus has the biggest number of possible inconsistent pairs with other states, its rate was relatively low (about 23 %). Also the state 11111 was valued consistently by 3/4 of the subjects.

The number of inconsistent states increased statistically significantly ($p<0.0001$) with age so that about 45% of respondents over 55 years had four or more states with inconsistencies (Table 3).

**Table 3. The percentage of respondents with inconsistent valuations by age groups (N=1272).**

| | Age groups | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Number of inconsistencies | 17-24 n=190 | 25-34 n=250 | 35-44 n=212 | 45-54 n=150 | 55-64 n=111 | 65-74 n=255 | 75-95 n=104 | All |
| None | 30.0 | 35.2 | 34.9 | 24.0 | 17.1 | 18.8 | 13.5 | 26.4 |
| 1 | 25.8 | 30.8 | 26.9 | 26.0 | 18.0 | 9.8 | 12.5 | 22.0 |
| 2-3 | 32.1 | 22.4 | 23.1 | 26.7 | 18.9 | 20.8 | 30.8 | 24.5 |
| 4 or more | 12.1 | 11.6 | 15.1 | 23.3 | 45.9 | 50.6 | 43.3 | 27.0 |

The mean VAS values of all included respondents (n=1272) and in the four inconsistency groups are seen in Table 4, where the states have been ranked in a descending order of values from the "no inconsistencies" group. The exclusion of respondents changed the absolute mean values of the states on average by 1.44 units (see second columns in Tables 2 and 4). Due to exclusion the mean values increased in 8 mild health states. In severe states the mean values decreased by about 2 units and in the state of being dead by about 3.5 units.

The differences between the mean values of the group with no inconsistencies and the group with one inconsistency were relatively small (Table 4). In 7 states the difference between the

mean values was greater than 5 units and it was statistically significant (t-test) in 9 states. The ranking order of the states did not differ significantly between these two groups (Figure 1).

The differences between the group with no inconsistencies and the group with 2-3 inconsistencies were statistically significant in 14 states. In 17 cases the difference was greater than 5 units and of them 14 states had a higher mean value in the group of 2-3 inconsistencies. Also the ranking order of states in the group of 2-3 inconsistencies was quite a bit different from that of the group with no inconsistencies (Figure 1).

**Table 4. The mean VAS values of the health states in groups with a different number of inconsistent valuations.**

| | Mean (all resp.) | S.D | n (N=1272) | No incons. | 1 incons. | 2-3 incons. | 4 or more incons. |
|---|---|---|---|---|---|---|---|
| 11111B | 91.20 | 13.90 | 1250 | 96.30 | 96.03 | 94.11* | 79.26* |
| 11111A | 90.80 | 13.97 | 1268 | 96.22 | 95.74 | 93.26* | 79.24* |
| 21111 | 77.53 | 17.41 | 707 | 79.82 | 79.83 | 75.46* | 74.97* |
| 11211 | 79.45 | 16.98 | 677 | 79.08 | 79.22 | 80.24 | 79.25 |
| 11121 | 75.76 | 17.33 | 677 | 75.54 | 73.75 | 74.61 | 78.88 |
| 12111 | 64.61 | 23.40 | 709 | 73.29 | 68.53* | 63.97* | 53.09* |
| 22111 | 70.51 | 18.01 | 227 | 69.50 | 68.44 | 72.80 | 72.17 |
| 12211 | 70.41 | 19.50 | 223 | 66.70 | 64.41 | 69.26 | 81.25* |
| 11112 | 64.92 | 21.92 | 704 | 63.71 | 63.01 | 60.98 | 70.99* |
| 12112 | 62.59 | 23.36 | 215 | 59.24 | 54.63 | 56.47 | 73.06* |
| 12121 | 54.99 | 23.90 | 212 | 58.36 | 57.43 | 50.92 | 54.56 |
| 22121 | 56.74 | 18.80 | 220 | 57.23 | 49.65* | 56.93 | 62.29 |
| 22211 | 55.60 | 21.15 | 216 | 55.97 | 54.15 | 54.49 | 57.03 |
| 11122 | 48.42 | 21.24 | 675 | 53.28 | 48.50* | 45.51* | 45.69* |
| 22112 | 53.28 | 18.52 | 217 | 52.49 | 54.82 | 48.32 | 58.86 |
| 12122 | 56.30 | 22.33 | 237 | 46.67 | 50.35 | 53.23 | 70.97* |
| 22221 | 45.94 | 22.70 | 214 | 45.34 | 41.46 | 45.32 | 49.35 |
| 11312 | 55.84 | 24.15 | 222 | 45.00 | 44.21 | 57.05* | 78.13* |
| 22212 | 50.65 | 20.64 | 220 | 44.72 | 43.98 | 47.72 | 66.31* |
| 22122 | 43.51 | 19.29 | 216 | 43.78 | 45.89 | 38.10 | 47.74 |
| 32211 | 46.59 | 23.84 | 702 | 40.34 | 42.97 | 43.31 | 58.54* |
| 12222 | 39.96 | 20.89 | 220 | 37.88 | 33.85 | 38.00 | 50.35* |
| 23321 | 37.59 | 21.12 | 218 | 37.49 | 36.22 | 34.23 | 42.85 |
| 21222 | 37.17 | 19.93 | 236 | 37.20 | 36.17 | 35.43 | 39.67 |
| 12223 | 46.69 | 27.06 | 212 | 37.13 | 33.27 | 38.71 | 65.35* |
| 21232 | 40.65 | 21.19 | 674 | 36.97 | 35.82 | 38.77 | 51.10* |
| 22222 | 36.62 | 19.90 | 219 | 33.49 | 32.05 | 35.30 | 45.37* |
| 31231 | 41.19 | 23.11 | 215 | 32.42 | 39.67* | 38.67 | 56.29* |
| 22331 | 41.42 | 23.73 | 222 | 31.94 | 32.61 | 38.23* | 65.55* |
| 11233 | 41.81 | 27.51 | 117 | 29.88 | 35.12 | 31.50 | 66.21* |
| 21133 | 38.54 | 26.36 | 210 | 29.74 | 25.93 | 29.13 | 56.32* |
| 23232 | 42.87 | 24.66 | 218 | 29.74 | 38.96* | 39.98* | 64.40* |
| 13332 | 41.12 | 27.04 | 221 | 27.71 | 30.56 | 41.63* | 66.64* |
| 13233 | 37.84 | 25.35 | 121 | 26.28 | 27.79 | 35.07 | 59.21* |
| 31323 | 37.86 | 26.11 | 215 | 25.33 | 32.02 | 36.60* | 58.63* |
| 32132 | 33.67 | 23.27 | 232 | 24.76 | 25.44 | 28.12 | 52.09* |
| 22233 | 34.16 | 23.50 | 667 | 23.36 | 26.35 | 31.48* | 56.21* |
| 22323 | 28.94 | 22.48 | 701 | 20.96 | 23.90 | 27.30* | 42.67* |
| 32232 | 27.94 | 22.71 | 224 | 19.81 | 18.61 | 25.10 | 48.42* |
| 32223 | 26.77 | 20.41 | 239 | 19.69 | 22.28 | 20.30 | 42.59* |
| 33321 | 24.94 | 23.14 | 671 | 18.52 | 21.27 | 21.91 | 38.79* |
| 32313 | 25.82 | 24.57 | 216 | 17.92 | 19.62 | 22.87 | 43.23* |
| 33122 | 24.01 | 23.20 | 216 | 17.71 | 15.16 | 22.26 | 41.11* |
| DeadB | 11.23 | 17.53 | 959 | 10.65 | 8.62* | 10.77 | 15.92* |
| DeadA | 11.23 | 17.51 | 957 | 10.50 | 8.50* | 10.71 | 16.39* |
| 33323 | 14.95 | 18.26 | 236 | 9.71 | 11.68 | 10.45 | 26.58* |
| 33333B | 15.51 | 23.74 | 1211 | 5.68 | 8.04* | 10.60* | 37.89* |
| 33333A | 15.58 | 24.09 | 1217 | 5.57 | 7.89* | 11.77* | 37.01* |
| Unconsc. | 12.35 | 24.92 | 677 | 5.50 | 7.18 | 6.32 | 30.89* |

* Difference to the group with no inconsistencies statistically significant ($p<0.05$)

The group with 4 or more inconsistencies produced a set of mean values that was quite different from the other three sets. Altogether 39 states had a significantly different mean value compared to the group with no inconsistencies. The increase in the number of inconsistencies decreases the mean values of the state 11111, does not affect much most of the mild and moderate states, and increases the values of the severe and very severe states excluding the state of being dead.

**Table 5. Single index OLS models for all respondents (N=1272) and different sub-groups of the inconsistent respondents together with model statistics.**

| Variable | Model 1 * All subjects | Model 2* No inc. | Model 3* 1 incons. | Model 4* 2-3 inc. | Model 5* 4 or more | Model 6* 0-3 inc. |
|---|---|---|---|---|---|---|
| 11111 | 91.0 | 96.3 | 95.9 | 93.7 | 79.2 | 95.3 |
| Constant | 10.7 | 12.6 | 14.9 | 14.7 | 2.3 (ns) | 13.9 |
| Mob 2 | 7.0 | 7.0 | 5.5 | 7.6 | 7.6 | 6.7 |
| Mob 3 | 18.8 | 21.2 | 18.4 | 20.4 | 14.9 | 20.0 |
| SC 2 | 8.4 | 10.0 | 10.1 | 8.9 | 5.1 | 9.8 |
| SC 3 | 10.4 | 10.3 | 10.4 | 11.5 | 9.8 | 10.7 |
| UA 2 | 4.7 | 8.6 | 6.8 | 3.6 | -0.54 (ns) | 6.4 |
| UA 3 | 10.9 | 17.0 | 13.9 | 8.8 | 2.2 (ns) | 13.5 |
| Pain 2 | 11.5 | 11.9 | 12.4 | 11.7 | 10.3 | 12.0 |
| Pain 3 | 11.7 | 15.2 | 15.1 | 12.2 | 3.8 ** | 14.3 |
| Mood 2 | 13.2 | 16.2 | 15.6 | 14.4 | 6.5 | 15.5 |
| Mood 3 | 13.0 | 16.0 | 15.9 | 15.1 | 7.0 | 15.6 |
| Dead | 69.1 | 73.1 | 72.4 | 68.2 | 60.7 | 71.4 |
| Uncons. | 67.9 | 78.2 | 73.8 | 72.6 | 45.9 | 75.1 |
| $R^2$ | 0.594 | 0.815 | 0.781 | 0.656 | 0.292 | 0.749 |
| $R^2$ adj. | 0.593 | 0.815 | 0.780 | 0.655 | 0.291 | 0.749 |
| F test | 2430 | 1994 | 1337 | 773 | 176 | 3683 |
| VIF test | 1.3-3.5 | 1.3-3.6 | 1.3-3.7 | 1.3-3.6 | 1.2-3.6 | 1.3-3.6 |
| BP test | P<0.001 | P<0.001 | P<0.001 | P<0.001 | P<0.001 | P<0.001 |
| Reset test | P<0.01 | P<0.01 | P<0.01 | P<0.01 | 0.05<P <0.01 | P<0.01 |

* parameters significant at the p<0.001 level if not indicated otherwise
** p=0.0095
*** p=0.0358
ns parameter statistically nonsignificant (p>0.05)

The modelling results are in Table 5 and 6. Model 1 is based on all included respondents (N=1272) and it produced consistent estimates for the levels of all dimensions except mood , where the difference between the levels 2 and 3 was reverse to what one would expect by 0.2

units (Table 5). Since the confidence intervals for the levels in mood dimensions were overlapping the parameter estimates are equal. The same assumption can also be made in all other models when the differences are below 1 unit.

The $R^2$ value of Model 1 was very good considering the individual level data that were used in the modelling. However in Model 2 that based on data from respondents without inconsistencies the explained variance in the data was about 80%. Also the explanatory power of Model 3 and Model 4 were very good and higher than that in Model 1. The data from respondents with 4 or more inconsistencies did not produce a good model (Model 5), since three parameters were not statistically significant at the 5% level and the $R^2$ value was low. Model 6 excluded the respondents with 4 or more inconsistencies. The $R^2$ value of the model was very high and the model produced consistent solutions for the tariff transformation as did Models 3 and 4. In all of these models the mood dimension and often also pain and self-care dimensions did not produce statistically significantly different parameter estimates for levels 2 and 3.

Valuewise the biggest health problem in all models appeared to be mobility level 3, followed by mood levels 2 and 3 (Table 5). Apart from Model 5 the parameter estimates were relatively close to each other in all models. The biggest deviation from Model 1 estimates were in the estimate of the upper anchor point (11111) and the constant. Apart from Model 5, all parameter estimates of the models were statistically highly significant ($p<0.001$).

The above models were also run with a data set, where all respondents without a value for the state of being dead were excluded (N=946). These two data sets produced nearly equivalent parameter estimates, e.g. in Model 6 the differences in the estimates were between 0 and 0.8. In addition these models were estimated by entering the N3 variable. However, the inclusion of the N3 did not improve the diagnostic statistics of these models. The models did not suffer from multicollinearity but they all had statistically significant problems with heteroscedastisity, and model specification error measured by Reset test (Table 5).

The log transformations of the individual data normalise the VAS scale values that are a little skewed downward. Model 7 shows the parameter estimates from the logit transformation data (Function 2) of the respondents with 0 - 3 inconsistent states (Table 6). The $R^2$ of Model 7 was the same as in Model 6 and also here all parameter estimates were statistically highly significant and consistent. Also in this model the mobility level 3 and mood dimension were valued as the biggest health problems. Apart from the previous models, the levels 2 and 3 were statistically significantly different in every dimension.

The threshold level where the inconsistencies do not influence significantly the modelling seems to be 3 inconsistent states. Because Model 6 produced the same parameter estimates for mood levels 2 and 3, the model was also estimated assuming equal value for these mood levels and adding N3 to model. Model 8 produced statistically significant value for N3 which

decreased by the level 3 estimates about 2 units in mobility, usual activities and pain dimensions (Table 6). The model statistics was basically the same as in Model 6. However, Reset test value of Model 8 (F=53; df=16030) was smaller than in Model 6 (F=87; df=16030) and the same as in Model 7.

**Table 6. Parameter estimates and tariff values for log transformation of VAS values together with tariff values for Model 6 when the states 11111 and being dead were the anchor points.**

| Variable | Model 7** logit (g1) | Model 8** (Model 6 with N3) | Tariff from Model 7 | Tariff from Model 8 |
|---|---|---|---|---|
| 11111 | 2.466 | 95.3 | 1.000 | 1.000 |
| Constant | 1.125 | 14.4 | 0.158 | 0.169 |
| Mob 2 | 0.263 | 6.0 | 0.058 | 0.070 |
| Mob 3 | 0.916 | 17.8 | 0.230 | 0.209 |
| SC 2 | 0.432 | 9.6 | 0.098 | 0.113 |
| SC 3 | 0.606 | 11.3 | 0.143 | 0.132 |
| UA 2 | 0.215 | 5.0 | 0.047 | 0.059 |
| UA 3 | 0.560 | 10.9 | 0.131 | 0.128 |
| Pain 2 | 0.482 | 11.4 | 0.111 | 0.134 |
| Pain 3 | 0.643 | 12.2 | 0.153 | 0.143 |
| Mood 2 | 0.671 | 15.2* | 0.160 | 0.178 |
| Mood 3 | 0.798 | 15.2* | 0.196 | 0.178 |
| Dead | 3.487 | 70.9 | 0.842 (=0) | 0.831 (=0) |
| Uncons. | 3.740 | 74.6 | 0.869 | 0.875 |
| N3 | - | 5.1 | - | 0.060 |
| R2 | 0.743 | 0.750 | | |
| R2 adj. | 0.743 | 0.750 | | |
| F test | 3571 | 3704 | | |
| VIF | 1.2 - 3.4 | 1.3-3.6 | | |
| BP test | P<0.001 | P<0.001 | | |
| Reset test | P<0.01 | P<0.01 | | |

* assumed to be the same

** all parameter estimates significant P<0.0001

The tariff values were calculated from Model 7 and Model 8 (Table 6). The tariffs produced scales that were relatively close to each other. It means that there is no difference which one is used in cost-utility analysis.

## 4 DISCUSSION

The response rate of the Finnish study was relatively high especially considering that the valuation task was quite difficult. For example in the US postal survey the response rate was 25.8% (Johnson et al. 1998). The high response rate may also have contributed to an increased number of blank, partly filled and inconsistently filled questionnaires. In this study the proportion of excluded respondents due to basic EQ-5D exclusion criteria was 22%, and if the missing value for the state of being dead was counted the rate increased to 42%. The proportion of exclusions is still relatively low considering the postal questionnaire used in the study. However, considerable proportion of the respondents did not value the state of being dead, which can lead to a big loss of data, if it is used as an exclusion criterion. Both genders and all age groups were well represented in the data and there were only few statistically significant differences between males and females in the background variables.

The big differences in the state-specific inconsistency rates (0.04 - 0.529) suggest that there may be some systematic factors that affect the number of inconsistencies per state. One plausible explanation is the location of the state on the page. Nearly all states with a very high rate located on the top of the page either on the left or right side of the page (e.g. states 11233, 12112, 12211, 12122, 12223, and 23232). Furthermore, many states which located in the right bottom corner just below state 33333 had a low inconsistency rate (e.g. 23321, 33122, and 33321). It means that many respondents have a tendency to draw a line from an upper corner state to a relatively high position on the VAS scale, especially if the mobility dimension is at the level of 1 or 2. Busschbach et al. (1997) have also found that the place of the state at the bottom of page, and in one case also at the top of page, produces a statistically significant framing effect compared to random order valuation of health states. The framing effects were on the same direction as in this study.

The exclusion of respondents that did not meet the EQ-5D inclusion criteria increased mean values of very mild states and decreased those of all other states. It shows that respondents that are uncertain over the valuation task have a tendency to assign relatively high valuations on the VAS scale. The same effect can also be seen when the mean values in different inconsistency groups are compared. The groups of no inconsistencies and with 1 inconsistency produced mean values that were very close to each other indicating that most of these inconsistencies were actually mistakes or mis-readings. Also the location of the health states on the two pages may have contributed to some of these inconsistencies in the groups of 1-3 inconsistencies.

The number of statistically significant differences in mean values in comparison to the group of no inconsistencies increased when moving from the group of 1 inconsistency to that of 2-3 inconsistencies. Also the preference ordering of the states in terms of mean values changed significantly especially in the moderate states. In addition the mean value of the states 11111,

33333 and some severe states (excluding being dead) changed. Since relatively few differences were statistically significant it seems that the group of 2-3 inconsistencies can be included in the modelling of data. However, respondents who had 4 or more inconsistencies did not produce logically consistent data for modelling.

The advantage of the used method to count inconsistencies was that it shows how many respondents has at least one inconsistency in the particular EQ-5D state. The method was a little different compared to the earlier EuroQol studies (Dolan and Kind 1996, Johnson et al. 1998). However, the rate of consistent answers is fully comparable between the studies. In the US postal VAS survey (Johnson et al. 1998) only 12% of respondents were consistent whereas in the interview based VAS valuation in the UK 57.4% of respondents were consistent (Gudex et al. 1996). It seems that the inconsistency rates of the postal surveys are much higher than in the interviews which may reduce their value as a method of getting valuation data in the future.

The single index models were made to all four inconsistency categories. The models showed that there were relatively small differences in the parameter estimates of the first three groups and that exclusion of the most inconsistent responses increased significantly the explained variance of the model including all the other respondents. The results suggest that the data also from other postal studies should be classified into different inconsistency groups and the mean values in these groups should be compared. This might reveal a threshold number of inconsistencies for excluding those respondents who seriously bias the results of modelling.

The production of the Finnish single index for the EQ-5D based on the data which included respondents having less than four states with inconsistencies. The mobility level 3 and mood levels 2 and 3 were the most important health factor for the Finnish respondents. In the Finnish data the levels 2 and 3 got values close to each other in self-care, pain and mood dimensions which can indicate that the level 2 problems are seen as more serious problems than e.g. in the TTO data in the UK (Dolan et al. 1996). The other more likely explanation is that the VAS valuation method produce downward skewed values and there is no place on the scale for variation in the lower end of the scale. It means that most of the respondents had not used the technique as having an interval property. The third possible explanation is that the Finnish translations have relatively small difference between the levels 2 and 3 on these dimensions. However, the very low percentages of level 3 health problems compared to level 2 health problems in the Finnish general population (Ohinmaa & Sintonen 1996) does not support the bias in translation.

N3 variable, that has got statistically significant values in most other single index models (Dolan et al. 1996, Johnson et al. 1998), does not get a significant value or get a low value in the Finnish data. The goodness of fit statistics of the models showed that the models excluding respondents with four or more inconsistent states explained a very high proportion of the variance in the data. However the models suffered from statistically significant

heteroscedasticity and specification error. The same problems have been found in all other models using individual level data. According to Dolan (1997), very little can be done to avoid these model specification problems in this valuation context. Since the goodness of fit of the models was very high, it is likely that these violations in the model assumptions do not have a significant effect on the estimated parameter estimates (Dolan 1997, Johnson et al. 1998).

The data used to model the Finnish tariff included respondents from 0 to 3 inconsistent states. The first tariff was made from the log transformed data. Since the method of transforming the model estimates back to 0 - 1 scale effect to the tariff values, there is some reservations in the use of this tariff. The other tariff model included N3 variable and the mood levels 2 and 3 were fixed to get the same parameter estimate. N3 got a statistically significant estimate but it was much smaller than in e.g. the UK, US or Spain. Although the fixed values for mood levels 2 and 3 reduce a little the ability of measuring change in longitudinal HRQOL studies, the inclusion of N3 variable increase the theoretical distance between levels 2 and 3. In the log based tariff only states 33333 (-0.011) and unconscious (-0.027) got a negative value. In the N3 tariff the value of state 33333 was -0.019 and there were four other states with negative values.

The study shows that the postal VAS technique can be used to elicit preference values for EuroQol with certain restrictions. First the postal method will produce significantly more inconsistencies than the VAS studies using interviews. That is why there should be some guidelines to exclude the most inconsistent respondents from the modelled data set. Secondly, the interval property of the VAS scale is not underlined in the postal survey, and it is possible that it produces a scale that is more skewed downward than the data collected by interviewers. The effects of the data collection method to the interval property of the VAS values can be analysed in the future, once the postal and interview based VAS data sets of the EQ-5D are combined to one data set.

The postal VAS valuation of the EQ-5D can produce a feasible data for modelling the single index for the measure. However the quality of the data is not as high as in the interview based data sets. The future VAS valuation data comparisons of different user countries could also include a comparison of the postal and interview based data sets. In addition, the comparisons between the TTO (and SG) and VAS methods, can produce important information for the choice of the valuation method in the future EQ-5D translations.

**ACKNOWLEDGEMENT:** The financial support from the Academy of Finland and Yrjö Jahnsson Foundation is gratefully acknowledged.

## REFERENCES

1. Abdalla M, Russell I. Tariffs for the EuroQol health states based on modelling the individual VAS and TTO data of the York Survey. In O'Hanlon M., Buxton M. (Editors), EuroQol Plenary Meeting, London October 1994, Conference Proceedings. Health Economics Research Group Research Report No. 20, September 1995, Brunel University, (1995) 75-92.

2. Badia X, Ohinmaa A, Roset M. Obtaining and modeling time trade-off values for EuroQol 5D health states. 1998; Submitted.

3. Busschbach J, McDonnell J, van Hout B. Testing different parametric relation between the EuroQol health description and health valuation in students. In Nord (ed), EuroQol Plenary Meeting, Oslo 17-18 Oct 1996, Conference Proceedings. Folkehelsa, Working paper No 2/1997, paper 12.

4. Dolan P. Modelling valuations for EuroQol health states. In O'Hanlon M, Buxton M. (Editors), EuroQol Plenary Meeting, London October 1994, Conference Proceedings. Health Economics Research Group Research Report No. 20, September 1995, Brunel University, (1995) 41-74.

5. Dolan P. Modeling valuations for EuroQol health states. Medical Care 1997;35:1095-1108.

6. Dolan P, Gudex C, Kind P, Williams A. The time-trade off method: results from a general population study. Health Economics 1996;5:141-154.

7. Dolan P, Kind P. Inconsistency and health state valuations. Social Science and Medicine 1996:42; 609-615.

8. Gudex C, Dolan P, Kind P, Williams A. Health state valuations from the general public using the visual analogy scale. Quality of Life Research 1996;5:521-531.

9. Johnson JA, Ergo A, and Coons SJ. Valuation of the EuroQol (EQ-5D) in adult US sample. PharmacoEconomics 1998;13:421-433.

10. Kmenta J. Elements of Econometrics. 2nd ed., New York. Macmillan Publishing Co. 1990.

11. Ohinmaa A, Helala E, Sintonen H. Modeling EuroQol values of Finnish adult Population. In: Badia X, Herdman M & Segura A (eds) Conference proceedings, The 12th EuroQol plenary meeting in Barcelona, October 1995, 1996:67-76.

12. Ohinmaa A, and Sintonen H. Quality of life of the Finnish population as measured by the EuroQol. In: Badia X, Herdman M & Segura A (eds) Conference proceedings, The 12th EuroQol plenary meeting in Barcelona, October 1995, 1996:161-172.

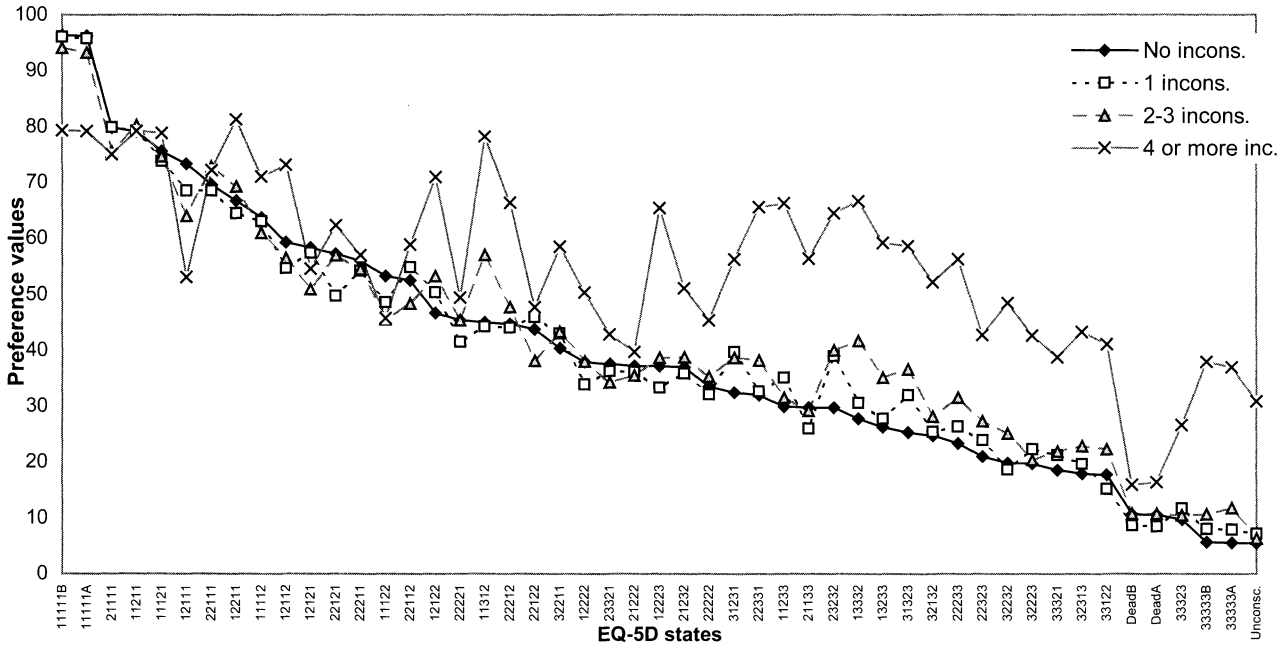


Figure 1. Mean values in the four inconsistency groups showed by decreasing values in no inconsistencies group.