University of Southern Denmark · Syddansk Universitet

Modelling Danish Weights for the 15D Quality of Life Questionnaire by Applying Multi-Attribute Utility Theory (MAUT)

Wittrup-Jensen KU[1,3] & Pedersen KM[2]

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

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

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

Health EconomicsPapers
2008:7

UNIVERSITY OF SOUTHERN DENMARK

Abstract

Background: The 15D is a multi-dimensional, standardised and self-administered instrument for measuring Health-Related Quality of Life (HRQoL). It is available in more than ten different languages including a Danish version. It can be used both as a profile and as a single cardinal index to estimate Quality Adjusted Life-Years (QALYs). At present only Finnish tariffs (weights) exist. However, as preferences for health may be country-specific, each country should develop their own set of 15D tariffs for use in economic evaluations.

Objectives: The objective is to estimate a set of national Danish 15D weights, which can be used to estimate QALYs and, e.g., apply these in economic evaluations. The aim is to explore alternative valuation techniques (additive and multiplicative) based on Multi-Attribute Utility Theory (MAUT). Furthermore, the aim is to assess validity in the alternative models, and finally to compare the results with the Finnish 15D weights.

Data and methods: All valuation tasks were carried out with self-administered postal questionnaires with no reminders. Five random samples (n = 1000 each) of the Danish non-institu-tionalised population aged 18 to 75 years were drawn from the Danish National Popula-tion Register. 1,260 out of the 5,000 questionnaires were returned, of which 57 were blank, a return rate of 24.1 per cent. In total, six models were estimated, two additive and four multiplicative, including models allowing health states to be valued as worse than death, i.e. negative va-lues. In order to assess the estimated models, focus was on feasibility, repeatability, validity (con-tent and construct), and the correlation (Pearson and Spearman correlation coefficients) between data from the models. Predicted valuations were also compared with the results from a Visual Analogue Scale (VAS) valuation task. Finally, the focus was on whether the models differed from the models estimated in the Finnish study by assessing the correlation between the two sets of weights.

Results: The study shows that it is possible to derive reliable and valid valuations for the 15D by applying the MAUT method in a postal-based questionnaire. It is possible to estimate an algorithm for all six models. However, the multiplicative models are all found to be inappropriate in estimating a single index for the 15D. Based on the best fitting model, the correlation between the Danish and the Finnish models is 0.98 ($R^2 = 0.97$) and significant ($p < 0.01$).

Conclusion: The two additive models are found to be appropriate in estimating a single index score for the 15D. It is recommended that model (2) should be used since it is more appropriate in describing the valuation of multi-attribute utility theory. Even though the correlation between Danish and Finnish weights for the 15D is high, Danish 15D weights should be applied in Danish HRQoL studies.

Introduction

The study reported here, to a considerable extent, replicates the Finnish study conducted by Sintonen (1994) in order to calculate a single index score for the 15D to be used to estimate QALYs and to apply these in the estimation of cost-utility ratios for various diseases. Currently, Finnish weights are applied in Danish CUAs because no Danish 15D weights exist. However, as preferences are likely to vary between the Finnish and Danish populations Danish weights ought to be applied in the estimation of QALYs and consequently in the estimation of cost-utility ratios. This would make the results more valid and relevant from a decision-making point of view regarding the application of health economic evidence in the prioritisation of scarce health care resources.

At present there is considerable uncertainty concerning the use of weights (or tariffs) developed in one country in analyses carried out in other countries. Is it acceptable, that is, are the preference structures sufficiently similar across country borders to allow the use of, say, UK weights in the rest of Europe? If yes/no, what are the consequences? Even for an established instrument like the EuroQol (EQ-5D), where one of the original aims was to compare preference values for health states across countries [The EuroQol Group 1990], it is rare to see cross-country comparisons [Badia 2001; Wittrup-Jensen *et al.* 2001], and the readily available UK tariffs are used in many contexts outside the UK for lack of local tariffs. The present study hence makes a two-fold contribution. The first aim is to estimate a Danish 15D set of weights, and the second to address the cross-country comparison issue by comparing the Danish and Finnish results. The exact methodology for cross-country comparisons is at present underdeveloped.

The 15D includes, as the name indicates, 15 dimensions of health: mobility, vision, hearing, breathing, sleeping, eating, speech (communication), elimination, usual activities, mental function, discomfort and symptoms, depression, distress, vitality, and sexual activity [Sintonen 2001]. Each dimension is divided into five levels. Compared to other similar instruments, e.g., the EuroQol (EQ-5D), the 15D is the most comprehensive in terms of dimensions included.

As the 15D defines an enormous number of mutually exclusive 15-dimensional health states (5^{15}), it is impossible to apply the usual stated preference methods in order to calculate a single cardinal index for each health state. Put differently: the complexity (cognitive overload) and number of health state descriptions makes it impossible to span the valuation space adequately in a traditional survey setting.

Objectives

The main purpose of this paper is to estimate an algorithm incorporating weights for the 15D HRQoL questionnaire that will result in a single cardinal index score and which, subsequently, can be applied in the context of estimating QALYs. The method for eliciting these weights is based on multi-attribute utility theory. However, since there is no standard approach concerning how the model ought to be structured (additive or multiplicative, e.g.), six different models are estimated. In assessing the best fitting model, focus is on issues such as feasibility, logical inconsistency, reliability, validity (content and construct), correlation coefficients (both Pearson & Spearman) between the mean index scores generated by each of the six models, and the cardinal score derived using the VAS. Since this study is essentially a replication of the method by which the Finnish weights for the 15D questionnaire were originally estimated, this study offers a unique chance of comparing the two data sets and assessing the correlation between the models estimated in each country. The two studies are compared using correlation analysis.

Modelling using Multi-Attribute Utility Theory (MAUT) within the 15D

The multi-attribute utility (MAU) method is a feasible and applicable possibility when the HRQoL instrument contains a large number of health states [Keeney & Raiffa 1993]. The first model is a two-stage additive valuation method proposed by Sintonen (1981b):

$$u(x) = V^{i}_{HM1} = \sum_{j}^{n} K^{i}_{j}[w^{i}_{j}(x_{j})] \qquad (1)$$

where V_{HM1} is the social value of health state H for individual i as produced by model 1, K_j is a positive constant for the j'th dimension, where (j = 1, 2, …, n), representing the relative importance the individual attaches to the dimension under the assumption that $\sum K^{i}_{j} = 1$, and $w^{i}_{j}(x_j)$ is a numerical function of the j'th dimension, representing the relative value of the five levels included in the dimension (top level = 1 and being dead = 0).

Model (1) is the simplest alternative as it explicitly assumes that the dimensions are additively independent for valuation purposes, and that the importance weights apply over the whole range of levels [Keeney & Raiffa 1993 pp. 295]. In other words, the difference in the relative importance between any two dimensions remains at a constant level. However, from a purely intuitive point of view, it seems more plausible that relative importance may change as a function of levels. Model 2, which incorporates this aspect, could be of the following form:

$$u(x) = V^i_{HM2} = \sum_j^n [K^i_j(x_j)][w^i_j(x_j)] \quad (2)$$

where $K^i_j(x_j)$ is a set of positive constants for the j'th dimension, representing the relative importance of the dimension at the different levels for the individual i ($\sum K^i_j = 1$ at any level) and $w^i_j(x_j)$ is a numerical function on the j'th dimension, representing the relative value the individual i places on different levels of the dimension (top level = 1 and being dead = 0). As in model 1, model 2 explicitly assumes additive independence.

Values elicited by models (1) and (2) have so far been applied in all 15-D applications [Sintonen 1981; Sintonen & Pekurinen 1988]. However, within the documented literature other models have been applied in an empirical context. An example is a multiplicative (dis)utility model formulated by Torrance *et al.* (1982):

$$u(x) = u_{HM3} = 1 - S_{M3}u^*_{HM3} \quad (3)$$

where u_{HM3} is the social utility of health state H as produced by model (3), S is a scaling factor, u^*_{HM3} is the social disutility of health state H defined as follows:

$$u(x) = u^{*i}_{HM3} = (\frac{1}{k})[\prod_j^n (1 + kk^i_j[w^i_j(x_j)]] - 1) \quad (3a)$$

and $w^i_j(x_j)$ is a numerical function on the j'th dimension, representing the relative utility the individual i places on different levels of the dimension (top level = 1 and being dead = 0). The k^i_j values here resemble the weights in models (1) and (2) with the difference that they are not scaled to sum to 1. However, the parameter k is related to the interaction parameter k^i_j as follows:

$$if \sum_j^n k^i_j > 1, then -1 < k < 0 \text{ (dimensions are substitutes),} \quad (3b)$$

$$if \sum_j^n k^i_j = 1, then\ k = 0, \text{ and the additive model holds} \quad (3c)$$

$$if \sum_j^n k^i_j < 1, then\ k > 0 \text{ (dimensions are complements)} \quad (3d)$$

Cases 3b-3d can be distinguished in terms of the multivariate risk postures they represent [Richard 1975]. Case (3b) represents multivariate risk aversion, case (3c) multivariate risk neutrality, and case (3d) multivariate risk-seeking behaviour. The attributes in case (3b) can be characterised as 'substitutes', while those in (3d) are 'complements' [Keeney & Raiffa 1993]. The interpretation is straightforward; substitute dimensions are such that an improvement in one is relatively satisfying, while an improvement in two or more is not that much better. Conversely, with complementary dimensions, an improvement in any one dimension alone is not very useful, while a simultaneous improvement in several dimensions is much better.

Torrance *et al.* (1982), allowing health states to be negative, fitted a power curve to the data using disvalues (disvalue = 1 - value) and disutilities (disutility = 1 - utility) in order to transform values into utilities. Hence one has to distinguish carefully between value functions and utility functions in this literature[1]. For person-mean the fitted disutility-disvalue relation is $u^* = v^{*1.6}$. The fitted function can be re-expressed in utility-value terms as $u = 1 - (1 - v)^{1.6}$. Model (3), using this power transformation, is applied to the 15D descriptive system and tested here.

An alternative to model (3), without the utility conversion (the power transformation), has also been proposed by Torrance (1982). It is the multiplicative multi-attribute disutility function:

$$u(x) = u_{HM4} = 1 - Z_{M4} u^*_{HM4}, \tag{4}$$

where u_{HM4} is the social utility of health state H as produced by model (4), z is a scaling factor, u^*_{HM4} is the social disutility of health state H defined as follows:

$$u(x) = u^{*i}{}_{HM4} = (\frac{1}{k})[\prod_j^n (1 + kk_j^i [w_j^i (x_j)]] - 1) \tag{4a}$$

and $w^i_j(x_j)$ is a numerical function on the j'th dimension, representing the relative value the individual i places on different levels of the dimension (top level = 1 and being dead = 0).

[1] A value function is (at best) what one obtains from using a visual analogue scale such as the ones used in the present study. Formally: a measurable value function $v(.)$ represents the judgment that if the strength of a preference for consequence (attribute) w over consequence (attribute) x exceeds the strength of preference for consequence (attribute) y over consequence z, then $v(w) - v(x) > v(y) - v(z)$ for all w, x, y, z. The question is whether it is possible to establish a potential relationship between measurable value functions and utility functions. Torrance's work with a power function of the form $u = 1 - (1-v)^b$ has attracted the most attention and hence is followed here. Note that the exponent b should be estimated from a dataset containing VAS-valuations and some variant of standard gamble valuations, i.e. u is a von Neumann-Morgenstern utility function. B=1.6 is not generally what is found in empirical studies. For instance Robinson (2001) found values of 4.50 and 20.9, showing in the context of the particular article that the power transformation was not stable across contexts – hence

In 1992 yet another model was put forward by Torrance *et al.* (1992), which contained a modified model (3). By applying the standard gamble method to derive utilities they obtained an estimated utility conversion factor of $u = 1 - (1 - v)^{2.29}$. In this fitted function they did not allow health states to take on a negative utility. A model for 15D based on these features is referred to as model (5) and the corresponding value model as model (6).

To summarise: All models presented here produce a social value or utility for all possible health states generated by the 15D instrument within a range of 0 (worst) and 1 (best). Respondents with no problems in any of the fifteen dimensions are given the best possible health state value, namely 1.

Material and Methods

The valuation tasks

The valuation tasks are described briefly below and related to the models outlined above, i.e. which task corresponds to the various models. An important difference between the various tasks was the inclusion/exclusion of explicitly stating the time of duration of the health states to the respondents and varying this duration. As will be seen later, all respondents did not complete all tasks. Instead respondents were randomly allocated to complete a limited number of tasks.

Task I: In *model (1)* respondents were first asked to choose the most important dimension and give it the value of 100 on an adjacent ratio scale (a 'ruler' ranging from 0 to 100). An arrow-shaped box followed each description of the 15D dimensions and the respondents were then required to draw an arrow/line to the ruler. The ranking/ordering of the 15 descriptions was determined randomly, however once determined, it was the same for all respondents. The valuation task (task I) was introduced to the respondents as:

"Below there is a list of some statements about health. People have different opinions of how important these statements are, depending on how they perceive health. In this study we are interested in your opinion.
To begin with we ask you to assess which of the statements about health shown below ***is the most important statement****, assessed from a health care point of view, i.e. the statement that you would be willing to give up last. Please, draw a line from the box (□) to 100 on the thermometer. Now we would like you to assess the importance of all the remaining statements compared to the most important statement. If for example a statement in your opinion is half (½ or 50%) as important as the most important statement, you draw a line from the box to 50 on the thermometer. If a statement in your*

casting doubt on the attempt to find a mathematical-statistical relationship between value and (von Neumann-Morgenstern) utility functions.

opinion is not important at all, seen from a health care point of view, you draw a line from the box to 0. In order to avoid misunderstandings we ask you kindly to state the number in each box to where you draw the line on the thermometer. In your assessment you can use all numbers between 0 and 100 that you find correspond with the given statement. The lines can cross each other and two or more statements can be given the same value. "

After this exercise all remaining dimensions were placed on the scale in relation to the most important dimension. For example, the arrow pointing to 80: read 8/10 as important as the most important dimension (80 per cent of the importance of the most important dimension). In the subsequent calculations these values were divided by 100 to bring them to a 0 to 1 scale and then transformed to satisfy the condition $\sum_{j}^{n} K_j^i = 1$. The social importance weights (K_j) were formed by averaging the respondent weights over the whole sample.

Task II: The respondents were asked to give a value to the different *levels* of each dimension on an adjacent 0 – 100 ratio scale. The value of 100 should be given to the best/most desirable level on each dimension and the other levels should then be valued (located) on the scale in relation to the best level. In subsequent calculations these values were divided by 100 in order to attain individual level values, and social level values [$w_j^i(x_j)$] were calculated by averaging them over the sample of the respondents. The *duration* of the time spend in the health states was *not* defined. In task II the instructions read as:

"On the next seven pages you will be introduced to different statements about health. These are the same questions as you answered in the questionnaire, but now we are interested in something different. We kindly ask you to assess how desirable these statements are, compared to each other. Please read the following information very carefully as it applies to all of the following seven pages. In the following we ask you only to assess those statements that you see on the page in question.
From the box that represents the statement on the top, which is the best imaginable health state, we have drawn a line to 100 on the thermometer. We now ask you to assess the desirability for you to be in the remaining health states, compared to the most desirable health state. If you, for example, think that a given health state is half (½ or 50%) as desirable as the most desirable health state, you draw a line from the box to 50 on the thermometer. From the box, which represents the least desirable health state, you draw a line to 0.
The reason for the presence of "unconscious" and "death" is that we are interested in the desirability of the statements compared to those two states. In order to avoid misunderstandings, we urge you to write the number in each box to where the line is drawn (e.g. 50). In your assessment, you are welcome to use all numbers between 1 and 99. The lines are allowed to cross each other and two or more statements can be given the same value."

Tasks III, IV and V: Three further versions of task II with an identical format, but focusing on *duration* were used, (tasks III, IV and V). Task III was similar to task II, but the *duration* of the states was defined at *one year*. The wording was: '*Imagine that the states last for one year. What happens after that is not known and should not be taken into account'*. In task IV the duration was *one month*. In task V the duration was again *one year*.

Task VI: For *model (2)* the weights for the bottom (lowest) level of each dimension, i.e. level 5 for each of the 15 dimensions, were elicited with a format resembling that of the EuroQol instrument (EuroQol Group 1990). An example could be within the mobility dimension: 'bed-ridden and unable to move around'. As in task I the respondents were asked to locate the best dimension and value the remaining dimensions accordingly. Here again the *duration* of the health states was unspecified. The values obtained were divided by 100 and transformed to satisfy $\sum_{j}^{n} K_{jb}^{i} = 1$ (b refers to the lowest level of dimension j). Social weights (K_{jb}) were formed by averaging the individual weights over the complete sample of respondents. The social importance weights for the intermediate levels were extrapolated linearly from the social weights of the extreme ends in relation to the distance between level values obtained from task II.

"Below is a list with health states which people can find themselves in at a given point of time. People can have different opinions of how good or bad these health states are. Here we are interested in <u>your</u> opinion.

Please, draw a line from the box (□) in each health state to the thermometer, which shows how good or bad you think this health state is, compared to the best and worst imaginable health state that you can imagine yourself in. The best imaginable health state is marked as 100 and the worst as 0 on the thermometer. In order to avoid misunderstandings we ask you to write the number in each box to where the line is drawn (e.g. 50). In your assessment you can use all numbers between 0 and 100. The lines can cross each other and two or more statements can be given the same value."

Task VII: In another version of task VI, the *duration* was defined to be one year. For models (3) – (6) the solution suggested by Torrance *et al.* (1982) was applied. The $w_j^i(x_j)$ values on a 0 to 1 scale (top level = 1 and bottom level = 0) were obtained from tasks II and V, and converted into utilities by using the power function $u = 1 - (1 - v)^{1.6}$ for model (3) and the power function $u = 1 - (1 - v)^{2.29}$ for model (5). The values for level five of each of the fifteen dimensions from tasks VI and VII were used

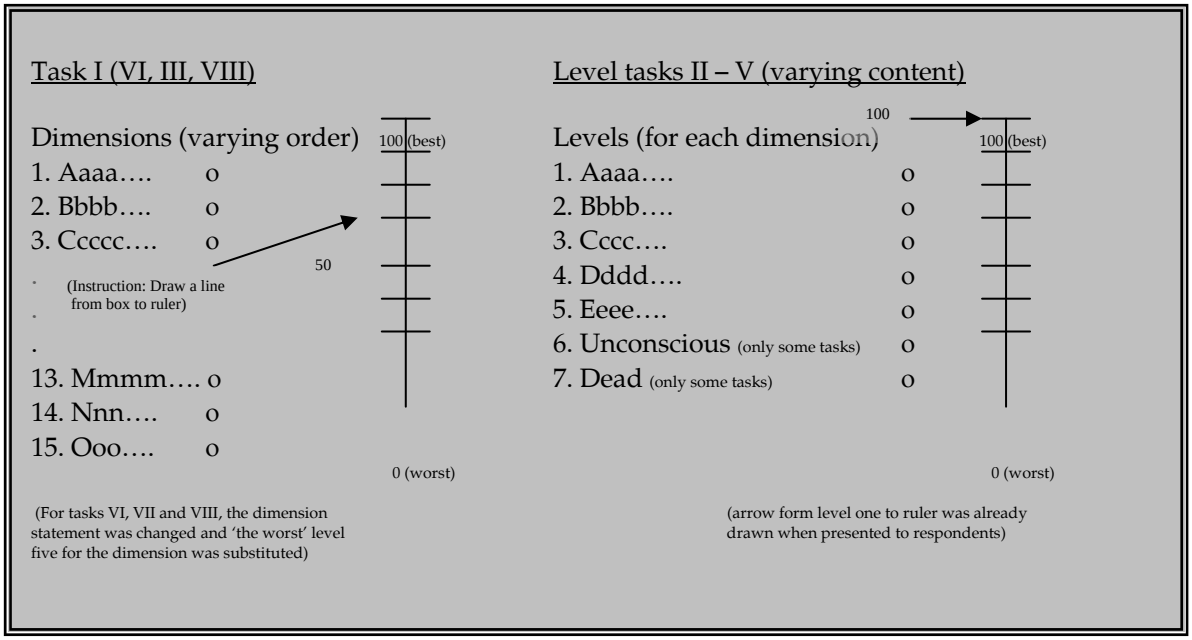
to derive the utilities for these corner states and the fifteen k_j values for models (3) and (5). (For models (4) and (6) the d_j values were applied).

Task VIII: For models (3) and (4) the value for the combination of the worst level (level five) for each dimension was derived by using a EuroQol-type format. The values were transformed to a 0 to 1 scale with the best imaginable health state being 1 and dead being 0. The value of the worst combination was then converted to a utility for model (3). The calculations required for models (3) – (6) are described in detail in Torrance *et al.* (1982) and (1992).

Task IX: Finally, a scale similar to that described in task VI was used asking the respondents to assess their own overall health status on the day they filled in the questionnaire (task IX).

A full description of how the models (1-6) and their algorithms were calculated is illustrated in appendix A.

Figure1. The general format of the valuation in the 15D exercise.



The above Figure shows the general format of the valuation exercise. The general principle was the use of a visual analogue scale for dimensions and levels separately. The exercise always started with the valuation of the 15 dimensions. Then the corresponding 5 levels for each dimension were valued on the ruler.

The samples

All valuation tasks (I to IX) were carried out as part of a postal-based questionnaire without any follow- up. Five samples, each randomised with regard to age, gender, and geographical domicile within

the Danish non-institutionalised population aged 18 to 75, were drawn from the Natio-nal Population Register. Each sample contained 1000 cases, which resulted in a total number of 5000 cases. The content of the questionnaire for each sample is shown in Figure 2.

Figure 2. Description of each sample.

Sample 1	Background data (age, gender, education, income, whether respondents experienced serious illness themselves, etc.), *tasks I, II and IX*, 15D questionnaire.
Sample 2	Background data, *tasks I, IV and IX*, 15D questionnaire.
Sample 3	Background data, *tasks VI, III and IX*, 15D questionnaire.
Sample 4	Background data, *tasks VII, V and IX*, 15D questionnaire.
Sample 5	Background data, *tasks VIII and IX*, 15D questionnaire.

At the end of the questionnaire all respondents were asked how long they needed to fill out the questionnaire (in minutes). The respondents were given around fourteen days to return the questionnaire in a pre-stamped envelope. 1,260 out of the 5,000 questionnaires were returned, of which 57 were blank. This was a return rate of 24.1 per cent, which is very low compared to similar postal-based surveys conducted in Denmark. We have not been able to identify the exact reason for the low response rate. No follow-up may be one explanation, but this does not explain the difference between the 24.1 per cent and the 50+ per cent in other similar surveys.

Of the 1,260 respondents, around 53 per cent were women, an over-representation of women compared to the general population, see table 1. There was also an over-representation of the age group 60-75 years, and under-representation of the age group 30-59 years. Normally, one would expect an under-representation of the elderly, especially when the tasks are quite complex as they were here. However, judging from the participation of the elderly, this standard hypothesis cannot be confirmed. We chose not to weight the sample, even though a difference between females and males was present, which one should be aware of when performing analyses where the gender factor plays a significant part. Performing gender weighting would not have had a significant impact on the results and, moreover, there could be adverse side effects such as an 'over-weighting' of cases, with error in measurement or other inaccuracies. Weighting was a possibility, but for operational reasons we chose not to do so.

Table 1. Representativeness in the 15D study judged by gender and age distributions.

	General population (18-75 yrs.) (N = 3,843,508) (January 1st 2000)	**15D study (N = 1,203) (Spring 2001)**

Gender:		
Male	50.1 %	47.0 %
Female	49.9 %	53.0 %
Age:		
18 – 29 years	21.8%	21.2 %
30 – 59 years	59.7 %	51.5 %
60 – 75 years	18.5 %	27.3 %

Evaluation criteria and theoretical evaluations

Following Sintonen (1994), the valuation methods and the resulting alternative value compo-nents were evaluated theoretically and empirically against four main criteria: feasibility, logical consistency, reliability, and validity. The theoretical evaluation criteria are presented briefly in this section while the empirical criteria appear in the following section.

Feasibility is judged empirically by measurement burden in terms of completion time and completion rates.

Consistency refers to the extent to which the respondents have a logical ordering of the health states [Dolan & Kind 1996; Badia *et al.* 1999]. Where a health state is logically better than another, its value should be higher. There is no logical order of importance between dimensions. However, the five levels within each dimension are clearly in a logical order of goodness. The consistency for each dimension is measured empirically by the percentage of respondents who assigned a set of values consistent with that order.

Reliability concerns the random variability associated with measurements. Ideally, this is a question of test-retest repeatability. However, this was not possible within the design used here. We focused on the repeatability and stability of valuations at the group or social level instead.

In the empirical context the repeatability of importance weights from the top of the scales was examined by comparing the results of an identical task I in samples 1 and 2, and from the bottom of the scales by comparing the results of task VI in sample 3 and task VII in sample 4. As already mentioned, the time duration specified in the two tasks varies. Hence it is expected that agreement may not be as good as that obtained with identical tasks for top-of-the-scale impor-tance weights. Pearson and Spearman correlation coefficients between the averaged sets of im-portance weights are the preferred statistical analyses. One-way analysis of variance is also applied and simple regression analysis is conducted. If the regression coefficient deviates from 1, the constant term from 0, or the fit is poor, the sets do not agree.

Validity indicates the extent to which accurate inferences about an underlying construct can be made based on a measure. As no gold standard exists for valuing health states, several types of validity have to be invoked.

Content validity relates to the adequacy of content of an instrument in terms of the number and scope of the individual questions that it contains, i.e. do these capture what they are supposed to capture. It makes use of the conceptual definition of the constructs being assessed, i.e. 'health', and consists of reviewing the instrument to ensure that it appears to be sensible and covers all of the relevant issues. Thus, content validation involves the critical examination of the basic structure of the instrument, a review of the development of the questionnaire, and also consideration of its applicability to the intended research question. With respect to the 15D, it is a good starting point to recall the aims of this instrument. The 15D was developed for use in several areas, but primarily for measuring the effectiveness of health care programmes in economic evaluation, that is, in cost-utility analysis [Sintonen 1994B].

As noted by Sintonen (1994A), it is a complex task to assess content validity. However, by initially checking whether the specification of duration makes a difference, it is possible in some way to obtain a sense of the presence/absence of content validity. Using the *Tukey multiple comparison tests* with one-way ANOVA, the existence of possible differences in mean valuations between samples with different duration specifications is tested.

Construct validation is one of the most important characteristics of a measurement instrument. It is an assessment of the degree to which an instrument measures the construct that it is designed to measure. The subject of construct validity is a difficult and controversial one. Validation involves gathering external empirical evidence, *convergent* or *discriminant*, so that meaningful inferences can be made from the measure. In order to show **convergent validity** the measure should correlate highly with other variables and other measures of the same construct, to which it should correlate on theoretical grounds. Furthermore, **discriminant validity** implies that the measure should not correlate with dissimilar, unrelated, variables or measures [Fayers & Machin 2000].

In order to assess convergent validity, the values produced by models (1) – (6) for the respondents' own health states were correlated (using Pearson correlation coefficients) with how respondents valued their own health in the VAS exercise, i.e. task IX, where the cardinal scores were transformed to a 0-1 scale.

The Danish weights were also compared with the Finnish weights. This was accomplished for the weights obtained by model (1). We looked at a one-way ANOVA between the two sets of weights, at

Pearson and Spearman correlation coefficients, and finally, a regression analysis was performed, followed by a scatter plot containing the best fitting line (as estimated by the regression analysis).

Results

Feasibility: Table 2 shows some statistics relating to feasibility. For convenience, the whole study is split-up into samples. In samples 1-3, which have comparable questionnaires, the response rate was 21.3 - 24.6 per cent. In sample 4, where the states 'unconscious' and 'dead' were not included for valuation, the response rate was 25.2 per cent. As sample 5 had only one valuation task, the response rate was clearly higher. The completion rates for importance weights were in the range of 77 - 88 per cent and were slightly lower for level values, especially for the state 'dead', underlining the well-known difficulties in valuing this state. The average completion time was in the range of 19 - 36 minutes as sample 5 only took, on average, 19 minutes to complete.

Table 2. Descriptive statistics relating to the samples and the feasibility of their tasks.

					Completion rates for levels (%)		
Sample no.	**Response rate (%)**	**Mean age**	**Male (%)**	**Completion rate for K_j (%)**	**Level 2**	**Being dead**	**Mean completion time (min.)**
1	21.3	47.4	49.3	88	70 – 79	67 – 71	32
2	22.0	45.7	46.6	87	71 – 79	68 – 71	33
3	24.6	45.8	40.7	86	77 – 83	75 – 79	36
4	25.2	47.4	50.2	87	77 – 86	NA	33
5	32.9	46.2	47.3	77	NA	NA	19

Logical consistency: The respondents in samples (1) - (4) valued the levels within each of the fifteen dimensions. As there is no way of telling what the values of the states 'unconscious' and 'being dead' were supposed to be, compared to the other levels, they were omitted. By focusing only on those respondents who filled out all levels within all fifteen dimensions, there were 14.0 - 27.6 per cent of the respondents who valued at least one level inconsistently. The dimension with the lowest percentage was 'vitality' and the highest 'speech'. When looking at each sample individually there were differences: The consistency percentage was higher in samples 3 and 4 compared with samples 1 and 2. The highest inconsistency across samples was in sample 4 within the levels in the dimension of 'speech' (31.9 per cent) and lowest in sample 2 within the levels in the dimension of 'usual activities' (9.7 per cent). However, no observations were excluded due to inconsistency.

Table 3. The mean K_j weights for the dimensions form different samples, and the final mean K_j weights from pooled samples.

Dimension	K_{j1} (top) Sample 1 (n = 190)	K_{j2} (top) Sample 2 (n = 192)	K_{jb3} (bottom) Sample 3 (n = 215)	K_{jb4} (bottom) Sample 4 (n = 222)	**Final K_j** Sample 1 + 2 (n = 382)	**Final K_{jb}** Sample 3 + 4 (n = 439)
Sleeping	0.0648	0.0653	0.0688	0.0680	0.0651	0.0684
Breathing	0.0741	0.0758	0.0615	0.0603	0.0750	0.0609
Eating	0.0714	0.0693	0.0536	0.0561	0.0703	0.0549
Speech	0.0702	0.0696	0.0688	0.0681	0.0699	0.0684
Mental function	0.0766	0.0770	0.0602	0.0636	0.0768	0.0620
Mobility	0.0658	0.0650	0.0503	0.0527	0.0654	0.0515
Discomfort/Symp.	0.0654	0.0652	0.0544	0.0583	0.0653	0.0564
Sexual activity	0.0561	0.0563	0.0825	0.0828	0.0562	0.0827
Hearing	0.0632	0.0631	0.0868	0.0775	0.0631	0.0820
Vitality	0.0732	0.0733	0.0726	0.0704	0.0732	0.0715
Distress	0.0598	0.0632	0.0727	0.0707	0.0615	0.0717
Usual activities	0.0672	0.0676	0.0708	0.0737	0.0674	0.0723
Elimination	0.0634	0.0623	0.0627	0.0608	0.0628	0.0617
Depression	0.0658	0.0673	0.0612	0.0637	0.0665	0.0625
Vision	0.0630	0.0599	0.0731	0.0734	0.0614	0.0733
Σ	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Reliability. The mean values for K_j weights for samples 1 to 4 are illustrated in Table 3. The one-way ANOVA showed that (pair-wise) none of the mean K_j weights between samples 1 and 2 differed significantly. The Pearson correlation coefficient between the two averaged sets of weights was 0.962 ($p < 0.01$) and the Spearman rank correlation coefficient was 0.953 ($p < 0.01$). When the sets of K_j weights of sample 1 were regressed on the sets from sample 2, the regression function was $K_{j1} = 0.004 + 0.939K_{j2}$ and R^2 was 0.93. The constant term did not deviate significantly from zero ($t = 0.82$, df 13) and the regression coefficient did not deviate significantly from 1 ($t = 12.78$, $df = 13$). Together these analyses indicate that the two sets agreed quite well, which means that reliability at the group level was acceptable. In order to obtain weights for the top of the scale dimensions samples 1 and 2 were pooled.

Looking at the reliability between samples 3 and 4, there were no significant differences (judged by pair-wise comparisons) between the mean K_{jb} weights as judged by the one-way ANOVA. The Pearson correlation coefficient was 0.96 ($p < 0.01$) and the Spearman rank correlation coeffi-cient was 0.95 ($p < 0.01$). When the set from sample 4 was regressed on the set from sample 4, the regression function was $K_{jb3} = -0.012 + 1.188K_{jb4}$, $R^2 = 0.92$. The constant term was not significantly different from zero ($t = -1.91$, $df = 13$) and the regression coefficient not significantly different from 1 ($t = 12.18$, df = 13). Overall the two samples agreed quite well and the reliability at the group level was good.

Validity. Regarding *content validity*, sample 3 (i.e. task III) and sample 4 (i.e. task VII) contained the same task with different durations - unspecified and one year, respectively. The one-way ANOVA resulted in no significant differences between the observed values of the two tasks, that is duration did not matter.

Turning to the issues surrounding *construct validity* it was our aim to compare all six models and the results of the respondents' valuations of their own health status by using the VAS exercise (task IX). By using the respondents' health status as measured by the 15D profile (the descriptive part), we were able to obtain mean values by applying models (1) to (6). The results are shown in table 4.

Table 4. Descriptive statistics of the values for models (1) to (6) and VAS scores.

Variable	Mean (SD)	Median	Min	Max	N	% < 0.58	% < 0.00
V_{HM1}	0.9411 (0.0649)	0.9615	0.5753	1.0000	1,170	0.00	0.00
V_{HM2}	0.9413 (0.0646)	0.9614	0.5814	1.0000	1,170	0.00	0.00
U_{HM3}	0.6470 (0.3785)	0.7590	-0.9300	1.0000	1,170	28.31	6.72
V_{HM4}	0.4362 (0.4534)	0.5126	-0.5722	1.0000	1,170	54.23	19.54
U_{HM5}	0.8371 (0.1985)	0.9059	0.1005	1.0000	1,170	10.65	0.00
V_{HM6}	0.5541 (0.3124)	0.5439	0.1091	1.0000	1,170	51.67	0.00
VAS	0.8723 (0.1321)	0.9000	0.1000	1.0000	1,124	3.92	0.00

As can be seen in Table 4, the mean values between models (1) and (2) are very close to being identical. The mean for self-valued health on the VAS, however, was lower than the mean obtained for the two models. Worth also noting is that the minimum value for the VAS score was 0.10, which was very low compared with both models (1) and (2). In models (3) and (4) negative values were allowed, i.e. health states worse than death, which evidently resulted in lower mean values for these two models. The mean score for model (3) was around 0.65, where the lowest (individual) score was close to -1 and around 7 per cent of all respondents had scores worse than death (value 0). Model (4) showed similar results, where the mean score was around 0.44 and over 50 per cent of the respondents had a negative score. Models (5) and (6) are also multiplicative models and differed only from models (3) and (4) in the sense that the transformation equation is changed. This change resulted in a mean score for model (5) of around 0.84, with around 11 per cent of the respondents having a score below 0.58, which is the lowest score among the respondents when applying the scores from model (2). The mean score for model (6) was around 0.55, with over 50 per cent of the respondents having a score below 0.58.

Table 5 shows that the correlations (Pearson) between the seven sets of scores (including VAS) were significant ($p < 0.01$). Nevertheless, the correlation coefficients varied quite considerably - from around 0.53 to 1.000. There was a perfect correlation (indicated by a correlation coefficient of 1.000) between the two additive models (1) and (2). Even though the correlations between all six models and the VAS scale were significant, the correlation coefficients appeared quite low compared to the correlation coefficients between the six models. The VAS scores correlated with models (1) and (2) and were significant ($p < 0.01$). With some reservations one could say that these findings provide, at least to some degree, solid convergent evidence of construct validity for the 15D value components based on models (1) and (2).

Table 5. Correlations of model (1), model (2) and the VAS scores.

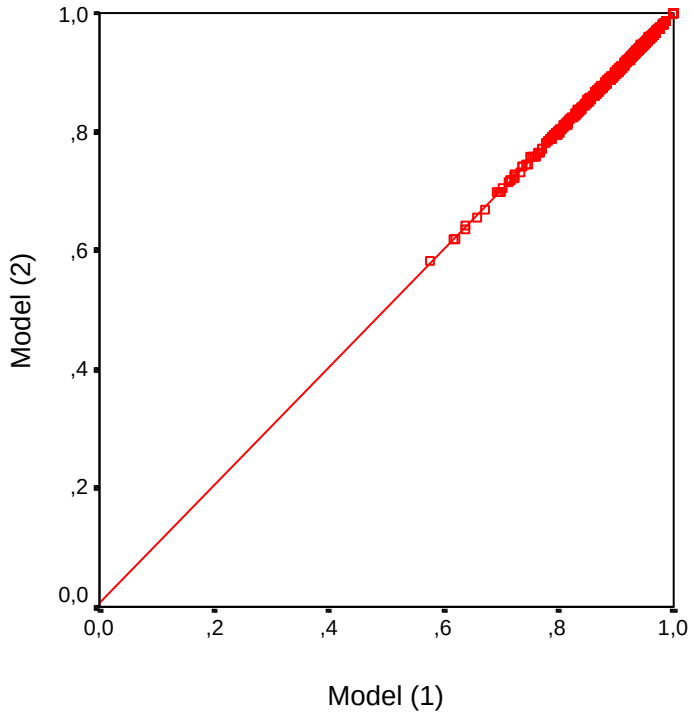
	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)	VAS
Model (1)	1.000						
Model (2)	1.000*	1.000					
Model (3)	0.979*	0.979*	1.000				
Model (4)	0.926*	0.926*	0.944*	1.000			
Model (5)	0.936*	0.934*	0.984*	0.893*	1.000		
Model (6)	0.841*	0.841*	0.867*	0.978*	0.810*	1.000	
VAS	0.665*	0.666*	0.658*	0.595*	0.647*	0.529*	1.000

*($p < 0.01$).

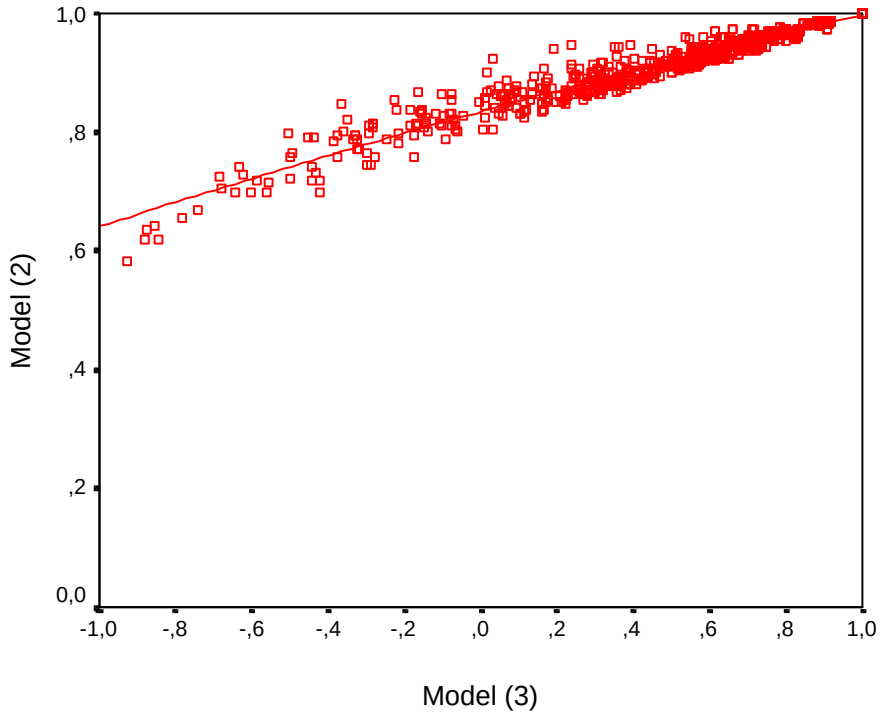
In addition to the correlation coefficients presented in Table 5 we also estimated the best regression equations for converting the scores from the other models into model (2). This was undertaken by plotting the models against model (2) and then finding the best fitting regression equation. Below we show both the plots and the equations.

Figure 2. Model (2) scores plotted against the remaining models and the VAS scores.

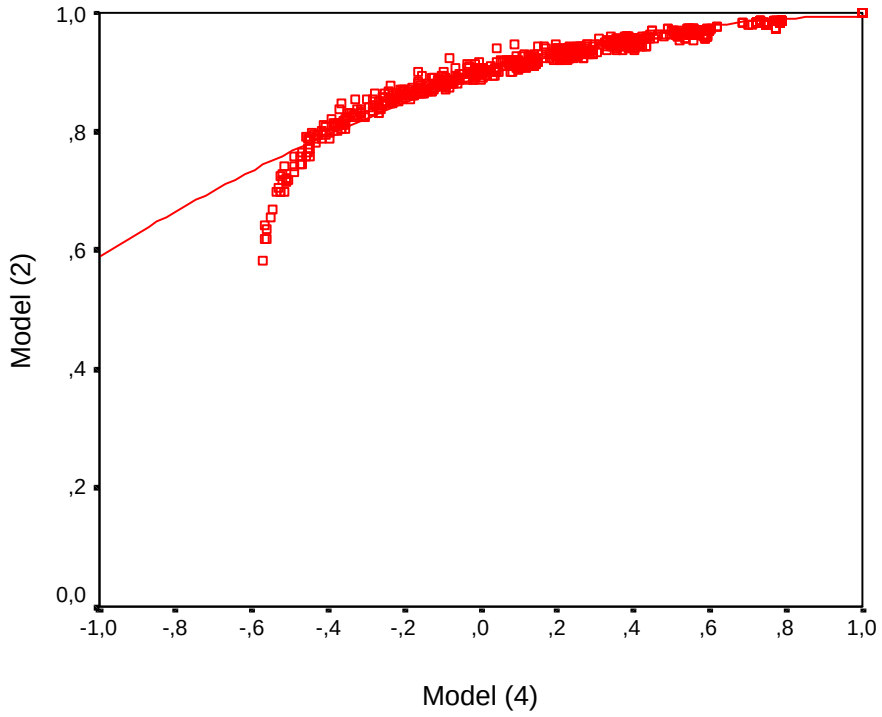
(a) Model (2) versus Model (1)



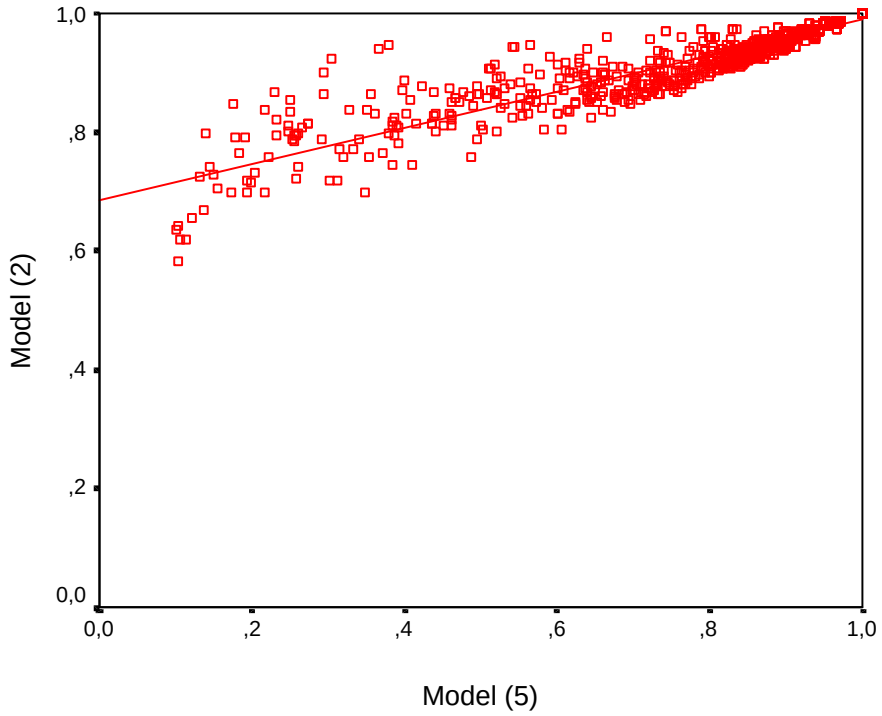
(b) Model (2) versus Model (3)



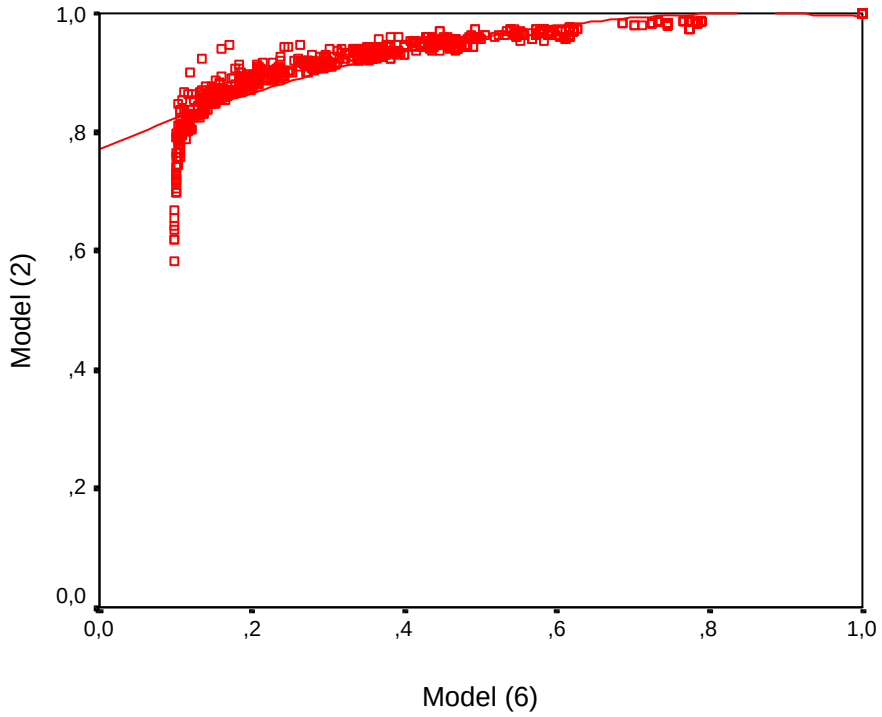
(c) Model (2) versus Model (4)



(d) Model (2) versus Model (5)



(e) Model (2) versus Model (6)



(f) Model (2) versus VAS scores

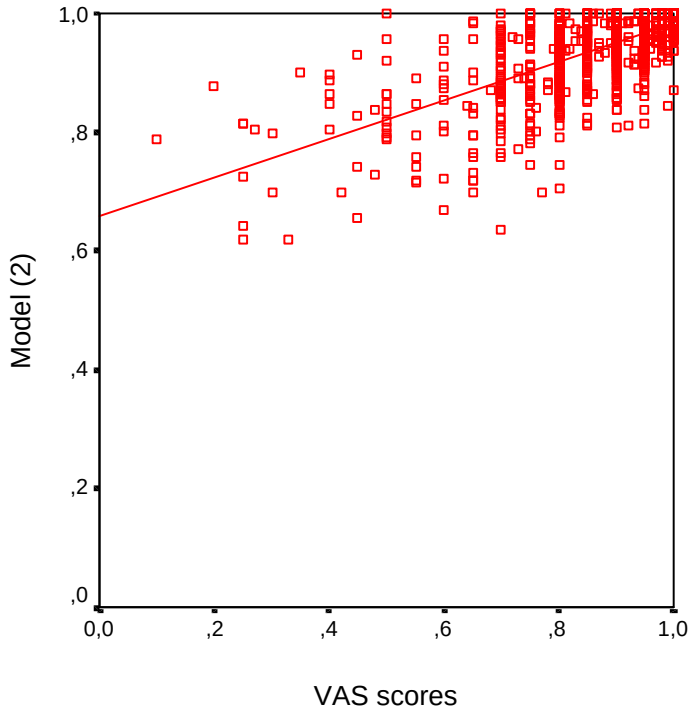


Figure 3. Best fitting regression equation for all five models compared with model (2).

V_{M2}	=	$0.00575 + 0.994*V_{HM1}$,	$R^2 = 1.000$
V_{M2}	=	$0.860 + 0.178*\ln(V_{HM3} + 1)$,	$R_2 = 0.898$
V_{M2}	=	$0.891 + 0.168*\ln(V_{HM4} + 1)$,	$R^2 = 0.949$
V_{M2}	=	$0.687 + 0.304*V_{HM5}$,	$R^2 = 0.934$
V_{M2}	=	$1.009 + 0.085*\ln(V_{HM6})$,	$R^2 = 0.873$
V_{M2}	=	$0.658 + 0.326*V_{VAS}$,	$R^2 = 0.666$

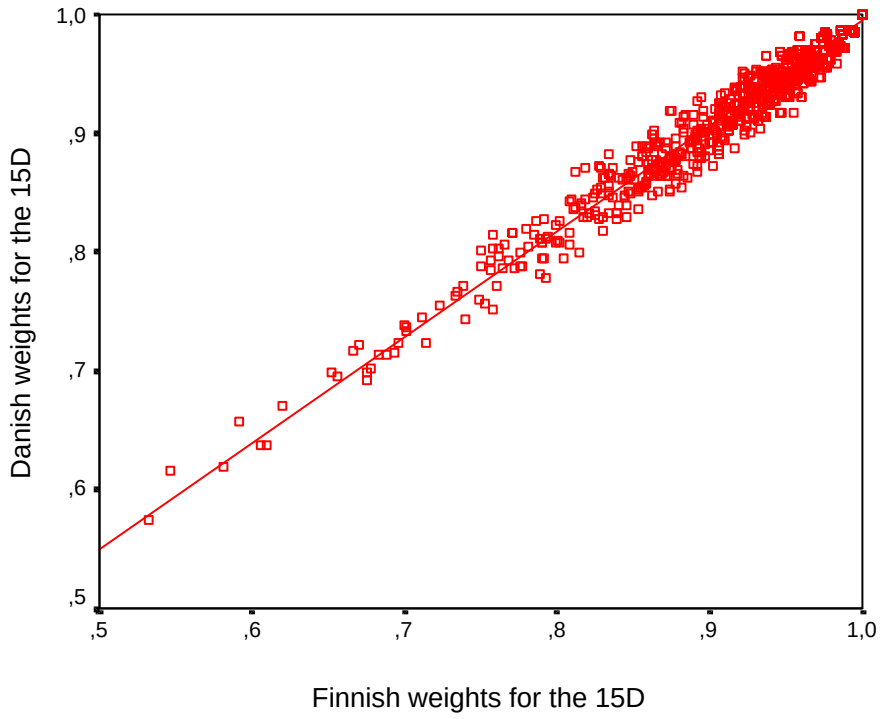
Comparison of Danish and Finnish 15D weights

We compared the Danish and Finnish scores for the sample of respondents within this study for model (2). The one-way ANOVA did not show any significant differences ($p < 0.01$). The Pearson correlation coefficient was 0.99 and significant ($p < 0.01$) and the Spearman correlation coefficient was 0.98 and also significant ($p < 0.01$). When the Danish weights were regressed on the Finnish weights we obtained the regression function: $W_D = 0.105 + 0.890W_F$, $R^2 = 0.97$. The constant term was not significantly different from zero ($t = 24.54$, $df = 1,168$) and the regression coefficient not significantly different from 1 ($t = 194.97$, $df = 1,168$). The scatter plot between the two weights, including the best fitting line as estimated by the regression analysis, is illustrated in Figure 4. It can be seen that the two sets of weights agreed quite well. However, more work has to be done, i.e. application in actual cost-

utility analyses, in order to say anything conclusive about whether the difference between the two sets really matters.

Although both the Pearson and Spearman correlation coefficients were high and significant, these tests do not give the whole picture. A head-to-head comparison between the Danish and Finnish weights for model 2 showed that there were differences both in the importance weights for dimensions as well as for the levels. Worth noticing is that the weights for level five differed considerably, as the Danish weights were significantly higher than the corresponding Finnish weights. For some reason, the Finnish population value for being at level five in one of the fifteen dimensions, was worse than that of the Danish population. The next step could be to look at the ranking (and/or numeric values) of the weights between the two sets.

Figure 4. Scatter plot between Danish and Finnish weights. Model (2).



Discussion

The present study has shown that it is possible to derive reliable and valid valuations for the 15D by a postal-based self-administered questionnaire. Hence, the Danish experience confirms the Finnish results.

However, the return rate of around 25 per cent was fairly low compared to normal Danish return rates and also lower than in the Finnish study, where the return rate was 43 – 46 per cent. It may be the case that the methods facing the respondents were too complex, but this was to some degree contradicted by the relatively high return rate from the elderly. Perhaps the tasks were simply too compli-

cated for a questionnaire in a postal-based study where there was no prior explanation. Alternatively, if a follow-up procedure had been used it would probably have boosted the return rate. It is unclear what is to be gained by using personal interviews, but it is an obvious alternative, albeit considerably more costly. The question is whether an inter-view-based study would improve both reliability and validity, since in this study both were at an acceptable level, but it undoubtedly would increase the participation rate.

A problem within the valuation of the health states is inconsistency. This is not a new phenomenon and is present in other applications of stated preference techniques, e.g. using the time trade-off method in estimating EQ-5D tariffs [Wittrup-Jensen et al. 2001]. However, the crucial question to be answered is what to do about this problem.

First, it should be recalled that consistency is an essential part of preference revelation, at least if we assume an underlying rational choice model, as is the case with all preference elicitation techniques. Hence, a high degree of inconsistency seriously calls into question the basic notions underlying exercises like the present. Additionally, do we include/exclude inconsistent choices in the data analysis, or do we find a threshold, for example only respondents with less than 3 inconsistencies are to be included in a study? From a strictly theoretical point of view, there are only two approaches: exclude inconsistencies, i.e. assume rationality in the sense of consistent ordering, or include them (totally or partially), basically implying that we are uncertain about the sense in which we think about preferences when applying preference revelation techniques. Since no practical guidelines exist this issue needs to be addressed in future research.

In this study for some dimensions the logical inconsistency rate was greater than 30 per cent, i.e. 30 per cent of the respondents displayed at least one inconsistency in valuing health states. The next issue would be to look at how these inconsistencies affected the estimated values for the models. This could be done simply by leaving out all respondents with inconsistent valuations and comparing the resulting weights with the weights obtained by including the weights. It could also be interesting to look at how many inconsistencies each respondent displayed and whether there were any relationships between the display of inconsistencies and socio-economic characteristics such as schooling/education and age. Results on these issues are forthcoming.

In total, six models were estimated and all the models proved feasible. Given that there is no *gold standard* to test the validity of the health state valuations for the 15D descriptive system, it is impossible to test explicitly, based on validity, which of the six models should be recommended as a future algorithm in estimating a single index score for Danish 15D valuations. However, as pointed at by Nord (1992), validity (in the form of criterion validity) could be tested by examining to what extent preference statements elicited in the 15D correspond with preferences that are directly elicited, that is, preferences elicited through scaling methods such as the SG, TTO etc. We believe that this is the wrong direction

in which to go, since this inevitably needs more assumptions to be made, for example that the scaling method has to be regarded as a gold standard. In stead, as was accomplished in the Finnish analysis, one should look carefully into the models and their implications [Sintonen 1994A].

We found that around 7 and 20 per cent of the respondents, when applying models (3) and (4) respectively, produced negative values, indicating that the individual would be in health states regarded as worse than dead. These numbers appear unrealistically high, given that the sample was elicited from the general Danish population. We believe that these properties make models (3) and (4) inappropriate in describing peoples' health status on a cardinal scale and consequently as a method to estimate QALYs, for example. Nevertheless, models (3) and (4) would still be appropriate as indicators for changes in HRQoL over a given time-span. However, in such a case there would inevitably be a considerable compression of values towards the lower end of the scale, which may put emphasis on the degree of validity.

As suggested by Sintonen (1994A), 15 dimensions in the multiplicative models are not appropriate. As in the Finnish study, we obtained values for the interaction parameters (c and d) very close to -1. The implications are that the $c_j u_j(x_j)$ factors would have to be on average at least 0.56 before their product with 15 dimensions exceeds 0. For example in model (5) this value is somewhere between levels 3 and 4 and in model (6) between levels 2 and 3. The implications are that a health state of middle level 3 would be regarded as equivalent to the state 'dead'. As also noted by Sintonen (1994A) this is not very plausible, but on the other hand it explains the very strong compression of values towards the lower end of the scale. The conclusion is that the multiplicative models are inappropriate in estimating a single index score for the 15D.

One of the remaining questions is: should one use model (1) or (2)? As noted by Sintonen (1994) both models are fairly easy to use in a computational setting. From a theoretical viewpoint, Sintonen (1994) suggests that the scores from model (2) should be applied since this model assumes that the importance weights across levels may vary. We agree that this is an important characteristic of a model and hence also recommend that the scores obtained from model (2) should be used.

The Danish weights estimated by model (2) were compared with the equivalent Finnish weights. Both the correlation coefficients estimated using the Pearson and Spearman approaches were around 0.96, indicating a high correlation between the two sets of weights. This may be an indication that preferences for health do not differ much between Finland and Denmark. Applying the MAU technique within a culturally very different country could, however, result in totally different weights. The comparison between Denmark and Finland obviously does not settle in any definitive way the question of universality of preferences for health.

Since the weights between the two countries are so highly correlated, does it then matter whether Finnish weights are applied in Danish economic evaluation or vice versa? It is hard to say anything conclusive, as one has to do a full cost-utility analysis in order to spot any impor-tant differences. However, using country-specific weights, *ceteris paribus*, makes the economic evaluation more valid as a decision-making tool in a national context. Therefore, even though the differences appear to be minor, we urge that Danish weights be used in Danish studies con-cerning HRQoL.

References

1. Badia X, Monserrat S, Roset M, Herdman M. Feasibility, validity and test-retest reliability of scaling methods for health states: The visual analogue scale and the time trade-off. *Quality of Life Research* 1999; 8: 303-310.
2. Badia, X, Monstserrat R, Hermand M, Kind P: A comparison of United Kingdom and Spanish General Population time trade-off Values for EQ-5D Health States. *Medical Decision Making* 2001; 21(1): 7-16.
3. Dolan P & Kind P. Inconsistency and health state valuations. *Social Science and Medicine* 1996; 4: 609-615.
4. The EuroQol Group. EuroQol: A new facility for the measurement of health-related quality of life. *Health Policy* 1990; 16: 199-208.
5. Fayers PM & Machin D. *Quality of Life*. Wiley: Chichester. 2000.
6. Feeny D, Furlong W, Boyle M et al. Multi-attribute health status classification systems: Health Utilities Index. *PharmacoEconomics* 1995; 6: 490-502.
7. Keeney RL & Raiffa H. *Decisions with multiple objectives. Preferences and value tradeoffs.* Cambridge University Press: New York. 1993.
8. Nord E. Methods for quality adjustment of life years. *Social Science & Medicine* 1992; 34: 559-569.
9. Robinson A, Loomes G, Jones-Lee M. Visual analog scales, standard gambles, and relative risk aversion. *Medical Decision Making 2001*; 21(1): 17-27.
10. Sintonen H. An approach to measuring and valuing health states. *Social Science and Medicine* 1981; 15: 55-65.
11. Sintonen H & Pekurinen M. *15D: A 15 dimensional measure of health*. Paper presented at the Health Economists' Study Group Meeting, Brunel University, July 18-20, 1988.
12. Sintonen H. *The 15D-measure of health-related quality of life. I. Reliability, validity and sensitivity of its health state descriptive system.* National centre for Health Program Evaluation, Working Paper 41, Melbourne. 1994B.
13. Sintonen H. *The 15D-measure of health-related quality of life. II feasibility, reliability and validity of its valuation system.* National Centre for Health Program Evaluation, Working Paper 42, Melbourne. 1994A.
14. Sintonen H. The 15D instrument of health-related quality of life: Properties and applications. *Annals of Medicine* 2001; 33(5): 328-336.
15. Stillwell WG, Seaver DA, Edwards W. A comparison of weight approximation techniques in multiattribute utility decision making. *Organ Behav Hum Perform* 1981; 28: 62-77.
16. Streiner DL & Norman GR. *Health measurement scales: A practical guide to their development and use.* Oxford University Press: Oxford. 1989.

17. Torrance GW, Boyle MH, Horwood SP. Application of multi-attribute theory to measure social preferences for health states. *Operations Research* 1982; 30: 1043-1069.
18. Torrance GW, Zhang Y, Feeny D, Furlong W, Barr R. *Multi-attribute preference functions for a com-prehensive health status classification system.* CHEPA Working Paper Series No. 92-18, McMaster University: Hamilton. 1992.
19. Torrance GW, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions. Health utilities index. *PharmacoEconomics* 1995; 6: 503-520.
20. Von Neumann & Morgenstern O. *Theory of games and economic behaviour*. Princeton University Press: Princeton, New York. 1944.
21. Von Winterfeldt & Edwards W. *Decision analysis and behavioural research*. Cambridge University Press: Cambridge. 1986.
22. Wittrup-Jensen K, Lauridsen J, Brooks R, Gudex C, Pedersen KM. *Estimating Danish EuroQol tariffs using the Time Trade-Off (TTO) and Visual Analogue Scale (VAS) methods*. Discussion paper presented at the 18th Plenary Meeting of the EuroQol Group, Copenhagen September 6-7. 2001.

Appendix A

Calculation method for the algorithms for models (1) – (6)

Model (1):

Model (1) was structured as an additive model where we used data from task I (sample 1 & 2) and task II (sample 1). Task I gave us the relative values for each of the 15 attributes compared to each other. These values were used to estimate the relative value for the best health state (level 1) in each attribute. The sum of the relative values for level 1, for all 15 attributes, summed to 1. In order to estimate the relative weights for the remaining four levels (levels 2-5) in each attribute, we applied the relative value from the best level in addition to the values from task II, where the latter were based on their mean values.

Model (2):

Model (2) was also structured as an additive model, where we again used data from task II (sample 1) and data from task VI (sample 3). The (mean) values (taken from task II) were the same as those used in model (1). Also, the relative weights for level 1 for all attributes were maintained, but now we used the values from task VI to estimate the relative weights for level 5 for all 15 attributes, however, without them summing to 1. The relative values for the remaining levels (i.e. levels 2-4) we found by applying linear extrapolation.

Models (3) and (4):

For these two multiplicative models, we used data from task II (sample 1) and task VIII (sample 5). In the Finnish study, Sintonen combined data from task VIII (sample 5) with data from task V (sample 4), however, the two tasks do not match since the respondents are given different information, concerning the duration of time spent in the health states, in the two exercises. Thus we believed that the two tasks were incompatible and chose only to use data from task VIII. While model (4) is explicitly defined as a value model, model (3) consists of (dis)utilities, where the values are converted by the formula $u=(1-v)^{1.6}$. Below we show how we estimated the weights that we used in the algorithm in models (3) and (4):

MOBILITY							
Mean value	v	**v** = 1 - v	**u**	Model 3	Model 4	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.5707	0.4076
0.7144	0.6800	0.3200	0.1615	0.9342	0.8174		
0.4717	0.4100	0.5900	0.4299	0.8248	0.6633		
0.2596	0.1800	0.8200	0.7280	0.7033	0.5321		
0.0974	0.0000	1.0000	1.0000	0.5924	0.4294		
VISION							
Mean value	v	**v** = 1 - v	**u**	Model 3	Model 4	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.5148	0.3456
0.7846	0.7400	0.2600	0.1159	0.9600	0.8662		
0.5429	0.4400	0.5600	0.3955	0.8633	0.7117		
0.3811	0.2400	0.7600	0.6446	0.7772	0.6088		
0.1855	0.0000	1.0000	1.0000	0.6544	0.4853		
HEARING							
Mean value	v	**v** = 1 - v	**u**	Model 3	Model 4	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.5295	0.3616
0.7734	0.7300	0.2700	0.1231	0.9555	0.8570		
0.5439	0.4500	0.5500	0.3842	0.8611	0.7088		
0.2969	0.1800	0.8200	0.7280	0.7368	0.5659		
0.1621	0.0000	1.0000	1.0000	0.6384	0.4706		
BREATHING							
Mean value	v	**v** = 1 - v	**u**	Model 3	Model 4	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.5351	0.3677
0.7345	0.7000	0.3000	0.1457	0.9464	0.8395		
0.5552	0.4800	0.5200	0.3512	0.8709	0.7218		
0.3220	0.2000	0.8000	0.6998	0.7427	0.5720		
0.1533	0.0000	1.0000	1.0000	0.6323	0.4650		
SLEEPING							
Mean value	v	**v** = 1 - v	**u**	Model 3	Model 4	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.5149	0.3457
0.7859	0.7400	0.2600	0.1159	0.9599	0.8661		
0.6228	0.5300	0.4700	0.2988	0.8967	0.7580		
0.4103	0.2700	0.7300	0.6044	0.7911	0.6242		
0.1853	0.0000	1.0000	1.0000	0.6543	0.4852		
EATING							
Mean value	v	**v** = 1 - v	**u**	Model 3	Model 4	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.5732	0.4104
0.6501	0.6200	0.3800	0.2126	0.9127	0.7822		
0.4071	0.3500	0.6500	0.5020	0.7940	0.6275		
0.2131	0.1300	0.8700	0.8003	0.6716	0.5014		
0.0931	0.0000	1.0000	1.0000	0.5896	0.4269		
SPEECH							
Mean value	v	**v** = 1 - v	**u**	Model 3	Model 4	$k_j(v)$	$k_j(u)$

1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.5222	0.3536
0.7021	0.6400	0.3600	0.1950	0.9310	0.8120		
0.4698	0.3600	0.6400	0.4897	0.8269	0.6658		
0.2912	0.1300	0.8700	0.8003	0.7171	0.5457		
0.1737	0.0000	1.0000	1.0000	0.6464	0.4779		
ELIMINATION							
Mean value	v	**v** = 1 - v	**u**	Model 3	Model 4	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.5380	0.3708
0.7382	0.7000	0.3000	0.1457	0.9460	0.8386		
0.4430	0.3400	0.6600	0.5144	0.8093	0.6450		
0.2682	0.1400	0.8600	0.7856	0.7087	0.5374		
0.1488	0.0000	1.0000	1.0000	0.6292	0.4621		
USUAL ACTIVITIES							
Mean value	v	**v** = 1 - v	**u**	Model 3	Model 4	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.5338	0.3663
0.7782	0.7400	0.2600	0.1159	0.9576	0.8612		
0.5401	0.4500	0.5500	0.3842	0.8593	0.7064		
0.3344	0.2000	0.8000	0.6998	0.7437	0.5730		
0.1554	0.0000	1.0000	1.0000	0.6337	0.4663		
MENTAL FUNCTION							
Mean value	v	**v** = 1 - v	**u**	Model 3	Model 4	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.5668	0.4031
0.6808	0.6500	0.3500	0.1864	0.9249	0.8016		
0.4417	0.3800	0.6200	0.4654	0.8124	0.6486		
0.2676	0.1900	0.8100	0.7138	0.7123	0.5409		
0.1032	0.0000	1.0000	1.0000	0.5969	0.4333		
DISCOMFORT AND SYMPTOMS							
Mean value	v	**v** = 1–v	**u**	Model 3	Model 4	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.5634	0.3993
0.7333	0.6900	0.3100	0.1535	0.9387	0.8254		
0.4494	0.3800	0.6200	0.4654	0.8142	0.6507		
0.2656	0.1800	0.8200	0.7280	0.7094	0.5381		
0.1086	0.0000	1.0000	1.0000	0.6007	0.4367		
DEPRESSION							
Mean value	v	**v** = 1–v	**u**	Model 3	Model 4	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.5280	0.3600
0.7344	0.6800	0.3200	0.1615	0.9419	0.8311		
0.5441	0.4400	0.5600	0.3955	0.8576	0.7043		
0.3166	0.1800	0.8200	0.7280	0.7380	0.5671		
0.1645	0.0000	1.0000	1.0000	0.6400	0.4721		

DISTRESS							
Mean value	v	**v** = 1–v	**u**	Model 3	Model 4	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.5076	0.3379
0.7835	0.7300	0.2700	0.1231	0.9584	0.8630		
0.5884	0.4900	0.5100	0.3405	0.8850	0.7411		
0.3599	0.2000	0.8000	0.6998	0.7636	0.5940		
0.1962	0.0000	1.0000	1.0000	0.6621	0.4925		
VITALITY							
Mean value	v	**v** = 1–v	**u**	Model 3	Model 4	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.4923	0.3218
0.7685	0.7100	0.2900	0.1380	0.9556	0.8572		
0.5964	0.4900	0.5100	0.3405	0.8904	0.7490		
0.3876	0.2100	0.7900	0.6858	0.7793	0.6111		
0.2210	0.0000	1.0000	1.0000	0.6782	0.5077		
SEXUAL ACTIVITY							
Mean value	V	**v** = 1 - v	**u**	Model 3	Model 4	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.5094	0.3399
0.7457	0.7100	0.2900	0.1380	0.9531	0.8523		
0.4910	0.3700	0.6300	0.4775	0.8377	0.6791		
0.2977	0.1400	0.8600	0.7856	0.7330	0.5620		
0.1939	0.0000	1.0000	1.0000	0.6601	0.4907		

The 'mean value' is the mean value across respondents for all levels for all attributes. These values are standardised by explicitly setting the upper and lower limits at 1 and 0, respectively. As we are interested in the disvalue we subtract the standardised value from 1. The corresponding "disutility" (u) to v, we found by using the formula $u = (1 - v)^{1.6}$. The values of the $k_j(v)$ factor were found by using the values from level 5 of each attribute, e.g. 0.0974 for 'mobility'. As the value of the worst combination of levels from task VIII (sample 5) was -0.5823 (after rescaling where 0 indicated death), this suggested a scaling factor of 1.5823 for model (4) and 2.0838 ($1.5823^{1.6}$) for model (3). The factor $k_j(v)$ for 'mobility' was then found by stating (1 - 0.097)/1.5823 = 0.5707. The factor $k_j(u)$ was found in the same way, however, the level 5 values had to be transformed by using the formula $u = (1 - v)^{1.6}$. Having estimated both the $k_j(v)$ and $k_j(u)$ factors for all 15 attributes, they both summed to a value > 1 ($\sum k_j > 1$), which meant that the k factor had to be somewhere in the interval from -1 to 0 (i.e. the dimensions were substitutes). We found the exact value of k (for both models) by applying the formula from Figure1, where we found k = -0.9999. Having all the data we needed, it was only a matter of simple calculation to find the values/disutilities for model (4)/model (3). For example: the disutility for level 2 in "mobility" for model (3) was found by stating 1 – (-0.9999)*0.4076*0.1615 = 0.9342. The corresponding value for model (4) was found by stating 1 – (-0.9999)*0.5707*0.3200 = 0.8174. By doing this on all levels for all attributes, we were finally able to find the algorithms for both models (3) and (4).

Models (5) and 6):

Models (5) and (6) were similar to models (3) and (4), but we now used a different transformation formula: $u = (1-v)^{2.29}$. Illustrated below is how we found the weights applied in the algorithms for models (5) and (6). The "mean value" is, of course, the same as before.

MOBILITY							
Mean value	v	**v** = 1 - v	**u**	Model 5	Model 6	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	1.0033	0.8796
0.7144	0.6800	0.3200	0.0736	0.9353	0.6790		
0.4717	0.4100	0.5900	0.2987	0.7373	0.4081		
0.2596	0.1800	0.8200	0.6348	0.4417	0.1774		
0.0974	0.0000	1.0000	1.0000	0.1205	-0.0032		
VISION							
Mean value	v	**v** = 1 - v	**u**	Model 5	Model 6	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.9050	0.6945
0.7846	0.7400	0.2600	0.0457	0.9682	0.7647		
0.5429	0.4400	0.5600	0.2651	0.8159	0.4933		
0.3811	0.2400	0.7600	0.5334	0.6296	0.3123		
0.1855	0.0000	1.0000	1.0000	0.3056	0.0951		
HEARING							
Mean value	v	**v** = 1 - v	**u**	Model 5	Model 6	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.9310	0.7411
0.7734	0.7300	0.2700	0.0499	0.9630	0.7487		
0.5439	0.4500	0.5500	0.2543	0.8115	0.4880		
0.2969	0.1800	0.8200	0.6348	0.5296	0.2367		
0.1621	0.0000	1.0000	1.0000	0.2590	0.0691		
BREATHING							
Mean value	v	**v** = 1 - v	**u**	Model 5	Model 6	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.9408	0.7590
0.7345	0.7000	0.3000	0.0635	0.9518	0.7178		
0.5552	0.4800	0.5200	0.2237	0.8302	0.5108		
0.3220	0.2000	0.8000	0.5999	0.5447	0.2474		
0.1533	0.0000	1.0000	1.0000	0.2411	0.0593		

SLEEPING							
Mean value	v	**v** = 1 - v	**u**	Model 5	Model 6	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.9052	0.6949
0.7859	0.7400	0.2600	0.0457	0.9682	0.7647		
0.6228	0.5300	0.4700	0.1775	0.8767	0.5746		
0.4103	0.2700	0.7300	0.4864	0.6620	0.3393		
0.1853	0.0000	1.0000	1.0000	0.3052	0.0949		
EATING							
Mean value	v	**v** = 1 - v	**u**	Model 5	Model 6	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	1.0077	0.8883
0.6501	0.6200	0.3800	0.1091	0.9031	0.6171		
0.4071	0.3500	0.6500	0.3729	0.6688	0.3451		
0.2131	0.1300	0.8700	0.7269	0.3543	0.1234		
0.0931	0.0000	1.0000	1.0000	0.1118	-0.0076		
SPEECH							
Mean value	v	**v** = 1 - v	**u**	Model 5	Model 6	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.9181	0.7178
0.7021	0.6400	0.3600	0.0964	0.9308	0.6695		
0.4698	0.3600	0.6400	0.3599	0.7417	0.4125		
0.2912	0.1300	0.8700	0.7269	0.4783	0.2013		
0.1737	0.0000	1.0000	1.0000	0.2823	0.0820		
ELIMINATION							
Mean value	v	**v** = 1 - v	**u**	Model 5	Model 6	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.9458	0.7683
0.7382	0.7000	0.3000	0.0635	0.9512	0.7163		
0.4430	0.3400	0.6600	0.3861	0.7034	0.3758		
0.2682	0.1400	0.8600	0.7079	0.4561	0.1867		
0.1488	0.0000	1.0000	1.0000	0.2318	0.0543		
USUAL ACTIVITIES							
Mean value	v	**v** = 1 - v	**u**	Model 5	Model 6	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.9384	0.7547
0.7782	0.7400	0.2600	0.0457	0.9655	0.7560		
0.5401	0.4500	0.5500	0.2543	0.8081	0.4839		
0.3344	0.2000	0.8000	0.5999	0.5473	0.2494		
0.1554	0.0000	1.0000	1.0000	0.2454	0.0617		
MENTAL FUNCTION							
Mean value	v	**v** = 1 - v	**u**	Model 5	Model 6	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.9964	0.8658
0.6808	0.6500	0.3500	0.0903	0.9218	0.6513		
0.4417	0.3800	0.6200	0.3346	0.7103	0.3823		
0.2676	0.1900	0.8100	0.6172	0.4657	0.1930		
0.1032	0.0000	1.0000	1.0000	0.1343	0.0037		

DISCOMFORT AND SYMPTOMS							
Mean value	v	**v** = 1 – v	**u**	Model 5	Model 6	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.9904	0.8539
0.7333	0.6900	0.3100	0.0684	0.9416	0.6930		
0.4494	0.3800	0.6200	0.3346	0.7143	0.3860		
0.2656	0.1800	0.8200	0.6348	0.4580	0.1880		
0.1086	0.0000	1.0000	1.0000	0.1462	0.0097		
DEPRESSION							
Mean value	v	**v** = 1 – v	**u**	Model 5	Model 6	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.9283	0.7362
0.7344	0.6800	0.3200	0.0736	0.9458	0.7030		
0.5441	0.4400	0.5600	0.2651	0.8049	0.4802		
0.3166	0.1800	0.8200	0.6348	0.5327	0.2389		
0.1645	0.0000	1.0000	1.0000	0.2639	0.0718		
DISTRESS							
Mean value	v	**v** = 1 – v	**u**	Model 5	Model 6	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.8923	0.6725
0.7835	0.7300	0.2700	0.0499	0.9665	0.7591		
0.5884	0.4900	0.5100	0.2140	0.8561	0.5450		
0.3599	0.2000	0.8000	0.5999	0.5966	0.2862		
0.1962	0.0000	1.0000	1.0000	0.3276	0.1078		
VITALITY							
Mean value	v	**v** = 1 – v	**u**	Model 5	Model 6	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.8656	0.6272
0.7685	0.7100	0.2900	0.0587	0.9632	0.7490		
0.5964	0.4900	0.5100	0.2140	0.8658	0.5586		
0.3876	0.2100	0.7900	0.5829	0.6345	0.3162		
0.2210	0.0000	1.0000	1.0000	0.3729	0.1345		
SEXUAL ACTIVITY							
Mean value	V	**v** = 1 - v	**u**	Model 5	Model 6	$k_j(v)$	$k_j(u)$
1.0000	1.0000	0.0000	0.0000	1.0000	1.0000	0.8957	0.6782
0.7457	0.7100	0.2900	0.0587	0.9602	0.7403		
0.4910	0.3700	0.6300	0.3471	0.7646	0.4358		
0.2977	0.1400	0.8600	0.7079	0.5199	0.2298		
0.1939	0.0000	1.0000	1.0000	0.3219	0.1044		

In models (5) and (6) the worst combination was assigned a value (utility) of 0.1, which was the same as the value of that state in model (2) (taken from task VI, sample 3). In order to estimate the factors $k_j(v)$ and $k_j(u)$, we had to undertake some intermediate calculations. We took the value of level 5 (scaled within the range of 0 – 1), subtracted the scaling factor, and divided the result by 0.9 (1 – 0.1). We did this for all 15 attributes. This intermediate factor we then subtracted from 1, which resulted in the value of $k_j(v)$. The factor $k_j(u)$ was found the same way, however, here we first transformed the (15) values (**v**) into utilities (**u**) by using the formula $u = (1 - v)^{2.29}$. Having all the necessary data, we

proceeded in the same way as in estimating the weights for models (3) and (4) and found the weights for models (5) and (6).