University of Southern Denmark · Syddansk Universitet

Estimation of a Preference-based Index Measure of Health for SF-12 by Applying Multi-Attribute Utility Theory (MAUT) – an experiment

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
2008:8

UNIVERSITY OF SOUTHERN DENMARK

Abstract

Background: The 12-item Short-Form (SF-12) questionnaire is a shortened version of the SF-36 multi-dimensional generic measure of health-related quality of life. It is widely used in clinical trials and outcome assessments due to its brevity and alleged psychometric comparability with the longer SF-36. There currently exists a preference-based single index for SF-36, and experiments with a similar index for the SF-12 have been undertaken. Unfortunately, the results are not yet available. These preliminary findings are based on a holistic design, where a limited number of health states are valued and the remaining health states found by regression analysis. Given the possible number of health states within the SF-12 (app. 18,000), the sensitivity of the predicted valuations is not very robust.

Objectives: The overall objective is to investigate whether a preference-based index for SF-12 can be estimated by applying the decomposed design of Multi-Attribute Utility Theory (MAUT) in the form of two additive models. Since no other algorithms are available it is not possible to test for (construct) validity. Consequently, we use the results from a Visual Analogue Scale (VAS) exercise as a gold standard, and compare the results from this exercise with the results from the two models, in order to assess the validity of the two models.

Data and methods: The study was conducted as a postal-based questionnaire with no reminder. Two random samples (n = 1,000 each) of the Danish non-institutionalised population aged 18 to 75 years were drawn from the National Danish Register. The respondents were given around fourteen days to return the questionnaire in a pre-stamped envelope. 230 out of the 2,000 questionnaires were returned of which 5 were blank. The multi-attribute utility theory method was assessed by applying category scaling and magnitude estimation. In total, two additive models were estimated.

Results: In order to develop an algorithm for the SF-12, it was necessary to exclude the item of 'general health'. By doing so, it was possible to apply the MAUT technique and estimate algorithms for the two additive models, which then can be used to estimate a single index score for use as an input in economic evaluation studies. The correlation analysis showed that the mean scores estimated in the two models were statistically significant with the mean scores from the VAS exercise.

Conclusion: This study was a very preliminary study that attempted to estimate a scoring algorithm for the SF-12 instrument. Given the structure of the valuation exercises and the lack of previous studies, we cannot test for validity. However, we believe that applying the MAUT technique as an alternative to other scaling techniques is more appropriate in eliciting health-state utilities for the SF-12 and that future studies should focus on the application of the MAUT technique.

Introduction

Health-Related Quality of life (HRQoL) measures are becoming increasingly important for evaluating the effectiveness of medical interventions and assessing the health of populations. Preference-based instruments, a subset of HRQoL measures, allow comparison of overall health status in populations and in clinical settings, and are suitable for economic analyses [Gold *et al.* 1996].

Given the increasing demand for health services, the interest in documenting the cost-effectiveness of health care interventions has led to the development and promotion of generic profile measures of health, such as the Short Form 36 (SF-36) Health Survey [Ware & Gandek 1998]. Although the SF-36 health survey has proved to be useful for a variety of purposes [Ware et al. 1993; Ware *et al.* 1994], it is too long for inclusion in some large-scale health measurements and for monitoring efforts [Ware *et al.* 1996]. Given these considerations, the study group behind the SF-36 shortened it and developed the SF-12. Since then, they have also developed a new questionnaire with even fewer questions, called the SF-8 (and earlier SF-6).[1]

Neither the SF-36 nor SF-12 was originally developed as a health preference instrument. Initially the SF-36 was developed purely as a profile health status measure. Given the increasing interest in generic HRQoL instruments, which encompass both a profile and a single index measure, researchers started looking into whether (and *how*) the SF-36 could be modelled for use in calculating quality-adjusted life years (QALY) in economic evaluations.

Today there is a fully developed algorithm, i.e. weights for the SF-36, called SF-6D [Brazier *et al.* 1998 and Brazier *et al.* 2002] and research on a similar algorithm for the SF-12 is currently in progress [Brazier & Roberts 2001[2]]. Other approaches concerning a single index score for the SF-36 and SF-12 have also been undertaken [Fryback *et al.* 1997; Shmueli 1999; Lundberg *et al.* 1999]. However, none of these studies has applied the MAUT approach. The majority used scaling methods, i.e. the standard gamble, time trade-off, or visual analogue scale as methods for the elicitation of a single index score. However, given the ordinal structure of the SF-12 and the limited number of questions therein, it should be possible to apply the MAUT method.

[1] See the web page www.sf-36.com.

Objectives

The main aim of this study is to apply the MAUT approach to the SF-12 in order to find weights and to develop an algorithm that can be used to estimate a single index score for the SF-12. Although there is no gold standard, construct validity is assessed by comparing the findings with results from the VAS exercise. Also, validity is assessed by performing a correlation analysis of the two results (Pearson and Spearman coefficients).

Applications of the SF-12 questionnaire

In contrast to index-based multi-dimensional descriptive systems, e.g. the EuroQol (EQ-5D), profile measures such as the SF-12 provide information on a selected number of scales or dimensions of health-related quality of life (HRQoL). The dimensions within the SF-12 are not preference weighted and are not combined, thus leaving an array of scale scores [Johnson and Coons 1998].

The SF-12 has scale scores for four of the eight health concepts in the SF-36 (physical functioning, role-physical, role-emotional and mental health) using two items for each concept. The remaining four health concepts (bodily pain, vitality, social functioning and general health) are each represented by a single item. Consequently, SF-12 can be used to create an eight-dimension profile, which approximates the SF-36 profile, although each of the single scale scores within SF-12 is estimated with less precision. Further, the SF-12 can be used to create a summary score form, referred to as the PCS-12 (physical component summary) and MSC-12 (mental component summary), which closely represent the summary scores of the SF-36. Again, however, the scores achieved on the PCS-12 and MCS-12 are less reliable than the SF-36 based summary scores, as they are based on fewer items and fewer defined levels of health. On the other hand, given the fact that confidence intervals for group means are more determined by sample size, the trade-off between the reduced respondent burden (approximately 2-3 minutes to complete) and the precision in measurement may be worthwhile for large group studies of general populations [Ware *et al.* 1996].

The development and structure of the SF-12

Although the 36-item short-form (SF-36) health survey has proved to be useful for a variety of purposes it is, according to Ware *et al.* (1996), too long for inclusion in some large-scale health measurement and monitoring efforts. The question is whether a shorter version is capable of yielding satisfactory results. Two developments have led to a strategy for constructing a shorter version of the SF-36 health survey. Physical and mental health factors have been found to account for 80 to 85 per cent of the reliable variance in the eight SF-36 scales in both patient and general populations [McHorney *et al.* 1993; Ware *et al.* 1994; Ware *et al.* 1995A]. Also, in cross-sectional and longitudinal tests, the PCS-36

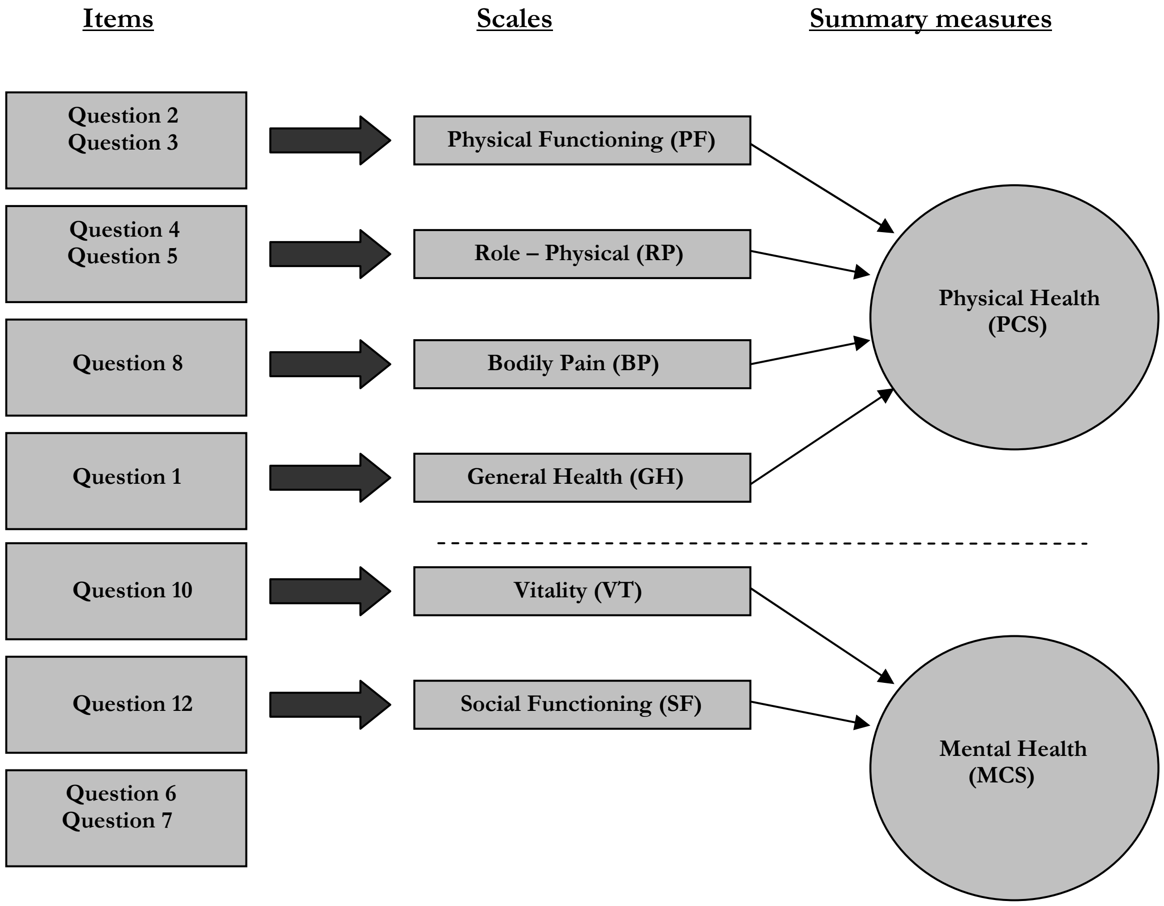
[2] Permission to quote this paper has been given by both authors.

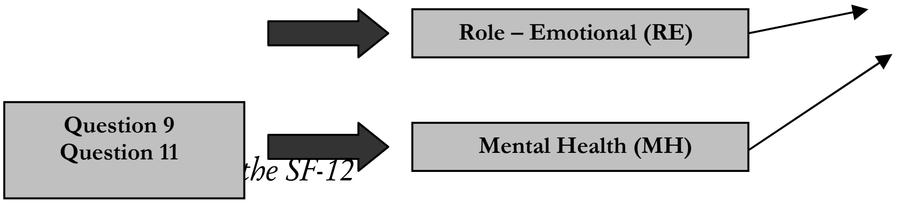
has shown hypothesized differences in nearly all tests based on physical criteria and the MCS-36 has detected hypothesized differences in all tests using mental criteria [Ware *et al.* 1994; Ware *et al.* 1995B].

According to Ware *et al.* (1996), the results observed for both the PCS-36 and MCS-36 measures indicate that it is possible to apply psychometric methods in order to reduce the number of health attributes assessed without substantial loss of information. The PCS and MCS measures also make it possible to construct an even shorter health survey, because the number of items in a survey is a function of the number of health attributes/dimensions for which separate scores are to be estimated with precision [Ware *et al.* 1995A; Nunally & Bernstein 1994]. The conclu-sion is that in cases where two summary scores (physical and mental health) are sufficient, it is possible to construct a shorter survey, which may prove to be valid and practical enough for more widespread applications.

The two SF-12 summary measures are constructed independently in order to reproduce the SF-36 physical and mental summary measures. Forward-step regression analysis was used to identify a subset of 12 or fewer items from the SF-36 and 2 weighting algorithms for estimating the PCS-36 and MCS-36. When choosing items Ware *et al.* (1996) looked at the representation of the eight SF-36 concepts and found that ten items were sufficient to reproduce both the PCS-36 and MCS-36 scores with an $R^2 > 0.90$.[3] They added two additional items in order to represent all eight concepts. Figure 1 is an illustration of the eight SF-12 health concepts.

Figure 1. Measurement model for the SF-12.





A the SF-12

An official Danish version of SF-12 exists, however, this is a translation of SF-12 version I. Since the publication of the SF-12 version I, the SF-36 Study Group has presented an up-dated version of the SF-12 called 'version II'. Although the changes are only minor, there is no official version in the Danish language of the SF-12 version II. Nevertheless, our aim is to estimate a set of tariffs based on the newest version of the SF-12. We therefore translated the English SF-12 version II into Danish. The changes in the English version II compared to version I were: 1) questions four to seven were changed from nominal scales, each with yes/no, into ordinal scales each containing 5 levels, and 2) the number of levels in questions nine to eleven were reduced from 6 to 5. With regard to the wording context, the short introduction now contains an example so that ho-pefully respondents better understand the terminology in the questionnaire. In addition the wording in questions 4 to 7 was altered to fit the changes due to the replacement of nominal with ordinal levels.

Eliciting valuations with SF-12 – past findings

To date only two studies have tried to investigate the relationship between the SF-12 and health-state utilities, based on the use of holistic scaling techniques such as the Time Trade-Off (TTO), the Rating Scale (RS), and the Standard Gamble (SG) [Lundberg *et al.* 1999; Brazier & Roberts 2001]. Given the paucity of studies, they are both summarised here.

The study by Lundberg *et al.* (1999) did not provide us with an algorithm that could be applied in our study. These analysts merely elicited health state utilities for the SF-12 using two scaling methods and compared these with each other using regression modelling. On the other hand, the objective of the study by Brazier and Roberts (2001) was to estimate an algorithm for the SF-12 that could be used in future studies to develop a single index score for all respondents. However, given that work was (and still is) in progress, we were not able to apply their algorithm to our sample results.

Applications using TTO and RS techniques for eliciting SF-12 utilities (Lundberg et al. 1999)

The data. This study was based on a self-administered postal questionnaire where health state utilities were measured using the RS (i.e. the VAS) and TTO methods. The VAS is a vertical, calibrated visual-analogue scale with labelled anchors of 'death' (at 0) and 'full health' (at 100). The respondents were asked to indicate with an arrow the point on this scale that illustrated their current health status. The 0-1 health-state utilities were obtained by dividing the number on the scale by 100.

[3] The eight concepts of the SF-36 are illustrated in Ware *et al.* (1996) pp 223.

The TTO question was phrased in the following way *"Imagine that you are told that you have 20 years left to live. In connection with this you are also told that you can choose to live these 20 years in your current health state, or that you can choose to give up some life years to live for a shorter period in full health. Indicate with a cross on the line below the number of years of full health that you think is of equal value to 20 years of full health that you current health state."* The 0-1 utilities with the TTO method were obtained by dividing this response by 20. In total 5178 respondents answered the VAS question and 4,740 answered the TTO question.

Methods. Lundberg *et al.* used linear regression modelling with the dimensions of the SF-12 entered as explanatory variables. Each item (except the intercept parameter) was entered as a categorical dummy variable, with the number of categories equalling the number of response alternatives. This implied that that they had to make no assumptions about the sizes of the utility differences between individual response alternatives. For each item, the worst level was used as the baseline category. The effect on health-state utility of every other category of a variable was estimated as the effect compared with that of the baseline category, and this was captured by the regression coefficient of each category. Finally, as a test for the stability of the estimated regression equations, the sample of each equation was randomly divided into two sub-samples of equal size. The regression equations were re-estimated for these sub-samples, and the results were examined to determine whether the estimated regression coefficients differed significantly between the sub-samples.

Results. Overall, the regression equations explained about 50 per cent of the variance in the RS values and about 25 per cent of the variance in the TTO values. The regression results for the RS were consistent and had significant effects for almost all response alternatives of the SF-12 items. The exception to this is the Role-Emotional Function (RE) dimension items that were not significantly related to the health state utilities. The results for the TTO were less strong than those for the RS. Many coefficients were not significant. In conclusion, most of the SF-12 items were related to the health-state utilities in the expected ways, with especially strong results for the RS method. The results obtained by Lundberg *et al.* suggest that the SF-12 could be conver-ted to health-state utilities, but as they put it themselves: *"... further work is needed to reliably estimate the conversion function."*

Applications using the SG technique for eliciting SF-12 utilities (Brazier and Roberts 2001)[4]

[4] Permission to quote from their findings granted by both authors.

The data. The study conducted by Brazier and Roberts was a comparison of the SF-12 and SF-6D, where the latter was used to elicit a single index score for respondents filling out the SF-36. The valuation study elicited 249 direct valuations defined by the SF-6D using the SG technique. The direct valuations were subsequently used to predict the remaining non-directly elicited valuations that could be described within the SF-6D (app. 18,000 health states) and within the SF-12 (app. 7,500 health states). They applied a split-sample design, where each respondent was asked to value six different health states. All respondents valued the worst health state - defined as 645655.[5] The remaining health states were classified as mild, moderate, or severe. All respondents with usable data were included in the study, including those respondents having missing values. In total 611 respondents, with up to six valuations each, entered the study (n = 3,518).

Methods. Both the SF-6D and SF-12 models were modelled at the mean level, which implied that the explanatory variables were used to estimate the mean value given to each of the 249 health states by the respondents who valued them. Brazier and Roberts also tested models that took account of individual variation across respondents (e.g. random effects models), which gave very similar coefficient estimates to the mean models, but were no better in terms of predictive ability (e.g. judged by R^2 values). A specific-to-general approach was used in order to derive a parsimonious regression model to estimate the mean value for each state. In addition, a binary dummy variable was included in order to take account of any additional effect on health state value, when one or more dimensions of health were at the 'most severe' level. In both cases the intercept in the regression model was restricted to equal unity.[6] The techniques applied in their valuation study were based on the assumption that health state 111111 (full health) equals one and death equals zero.

Results. On average, each of the 249 health states was valued around fifteen times. Mean health states valuations ranged from 0.10 to 0.99 and in general yielded large standard deviations. The relative ordering of health states within the SF-12 conformed to the logical ordering of the SF-6D. Furthermore, the majority of respondents valued the worst health state (645655) as being better than death. However, only a few health states were valued at full health (1.0), indicating the willingness of respondents to risk a worse health state in order to have the chance of a better state of health. In their conclusion Brazier and Roberts (2001) suggested that the SF-12 model was similar to the SF-36 model in terms of the coefficients on the SF-6D levels. The SF-12 was also slightly worse in terms of fit judged by explanatory power, mean absolute value of error, and percentage of predictions. Finally, there were no systematic differences between the SF-12 and SF-36 models, though both models suffered from a tendency to over-predict the lower end.

[5] See Brazier & Roberts (2001) for a review.

[6] In theory the intercept represents, as noted by Brazier and Roberts (2001), the value of full health, i.e. when each attribute of the health state is at level 1. However, empirical findings suggest that the value of the intercept is usually less than one [Dolan *et al.* 1997; Devlin *et al.* 2000; Wittrup-Jensen *et al.* 2001; Brazier *et al.* 2002].

Applications of Multi-attribute Utility Theory (MAUT)

Multi-Attribute Utility method is a feasible and applicable possibility when the HRQoL instrument contains a large number of health states [Feeny *et al.* 1995]. The first model is a two-stage additive valuation method proposed by Pliskin *et al.* (1980), Sintonen (1981), and Torrance *et al.* (1982). The two stages are concerned with the health dimensions, K_j, and the levels of each dimension, respectively:

$$u(x) = V^i_{HM1} = \sum_j^n K^i_j [w^i_j (x_j)] \tag{1}$$

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

Model (1) is the simplest alternative. It explicitly assumes that the dimensions are additively independent for valuation purposes, and that the importance weights apply over the whole range of levels. 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 would seem more plausible that the relative importance may change as a function of levels. Model 2, which incorporates this aspect, could be of the following form:

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

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 the individual *i* attaches to it ($\Sigma K^i_j = 1$ 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 death = 0). As in model 1, model 2 explicitly assumes additive independence. Values elicited by model (1) and (2) have so far been applied in all the estimations of a preference–index for the multi-attribute and preference-based HRQoL questionnaire 15D [Sintonen 1994].

Methods and data

The study

The study was undertaken as a randomly drawn postal-based questionnaire survey in May 2001 where 1,986 respondents were part of the initial sample. The age range was set at 18 to 75 years. The sample was drawn from the National Danish Register by using gender and domicile. The questionnaires mailed to the respondents contained a pre-stamped envelope, and they were given around fourteen days to return the questionnaire. The questionnaire included: 1) SF-12 questionnaire version II, 2) a Visual Analogue Scale (VAS) exercise, 3) a weighting exercise for SF-12 based on the MAUT approach and 4) background demographic questions.

The SF-12 questionnaire itself was based, as previously described, on an unofficial translation of the English version of SF-12 version II. The VAS instrument was an assessment of the respondents' own health at the present day, similar to the one contained in the EuroQol (EQ-5D) questionnaire [EuroQol Group 1990]. The endpoints were given by 0 (worst imaginable health state) and 100 (best imaginable health state). The weighting exercise contained two exercises using category scaling and one exercise using magnitude estimation. All respondents under-took two of the above exercises: the magnitude exercise and one of the category scaling exercises.

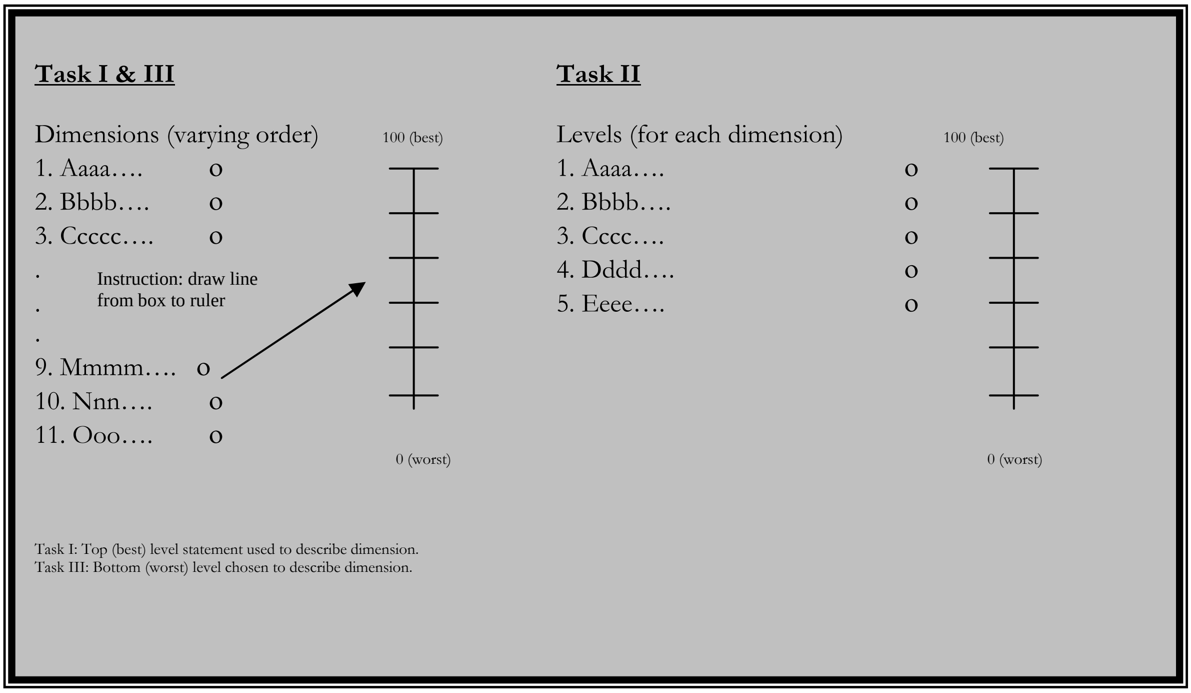
The background data collected were age and gender of the respondent, education level, persons in household, employment status, and income. In addition, the respondents were asked whether they used their own or other peoples' experiences while assessing the health states, and whether their assessment was influenced by something that could happen within the next ten years.

The valuation tasks

In *model (1)* the dimensions were described by their top levels (level 1) adjacent to a vertical continuous 0-100 ratio scale (task I). Each dimension description was followed by an arrow-shaped box. The order of descriptions was determined randomly, but was the same for all subjects. The values obtained were divided by 100 to bring them to a 0-1 scale and then transformed to satisfy $\Sigma_j K_{ij} = 1$. 'Social' importance weights (K_j) were formed by averaging the individual weights over the sample. At the second stage, the subjects were asked to give a value to the levels of each dimension, one dimension at a time, using the same format (task II). The levels were described adjacent to a 0-100 ratio scale of desirability. The *duration* of the states was not defined. The values obtained for each state were divided by 100, transformed linearly to meet $w_{ij}(x_{ju}) = 1$ (by definition for the top level). 'Social' level values $w_j(x_j)$ were calculated by averaging them over the sample.

For *model (2)* weights for the bottom level of each dimension were elicited using a format resembling that of the EQ-5D [EuroQol Group 1990] (task III). Here the *duration* of the states was again unspecified. The values obtained were divided by 100 and transformed to satisfy $\Sigma_j K_{ijb} = 1$ (*b* refers to the bottom level of j). 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.

Figure 2. General format of the valuation exercise in SF-12.


Task I & III
Dimensions (varying order)
1. Aaaa…. o
2. Bbbb…. o
3. Ccccc…. o
Instruction: draw line from box to ruler
9. Mmmm…. o
10. Nnn…. o
11. Ooo…. o
100 (best)
0 (worst)
Task II
Levels (for each dimension)
1. Aaaa…. o
2. Bbbb…. o
3. Cccc…. o
4. Dddd…. o
5. Eeee…. o
100 (best)
0 (worst)
Task I: Top (best) level statement used to describe dimension.
Task III: Bottom (worst) level chosen to describe dimension.

The samples

All valuation tasks were carried out using self-administered postal questionnaires with no reminder. Two random samples (n = 1,000 each) of the Danish non-institutionalised population aged 18 to 75 years were drawn from the National Population Register. The content of the questionnaire for each sample was:

Sample 1 The SF-12 questionnaire, the Visual Analogue Scale (VAS) assessment of own health, tasks I and II, and background data (age, gender, education, income, whether experienced serious illness etc.).

Sample 2 The SF-12 questionnaire, Visual Analogue Scale (VAS) assessment of own health, tasks III II, and background data (age, gender, education, income, whether experienced serious illness etc.).

Weighting dimensions?

In this valuation exercise we explicitly gave each attribute (dimension) in the SF-12 questionnaire a weight of 1. However in the case of the SF-12, as illustrated in Figure 1, each of the eleven attributes is combined into scales where four out of eight scales comprise *two* attributes. It is not *a priori* clear whether each attribute in these cases should have the same weight, i.e., in this case the weight of 1.

In what follows we discuss the pros and cons of whether or not there is evidence for weighting dimensions that are part of a larger scaling instrument. Given the limited evidence on explicit weighting in both HRQoL profile- and preference-based instruments, we turn towards more general Quality-of-Life (QoL) instruments. However, these findings and conclusions may be applicable to the area of both profile- and preference-based HRQoL instruments.

Although the notion of well-being is generally thought of as a global concept, it has been widely recognised that not all aspects of QoL are of equal importance to all individuals [Diener 1984]. For instance, individuals in good health may not consider matters of health as important as those in poorer health. It is no surprise that individuals vary in how much importance they place upon different domains. However, according to Trauer & McKinnon (2001), it is essential that one is aware how these domains have been selected, given that the criterion of importance may have been implicit in the selection.

QoL instruments that obtain both *Satisfaction* (S) and *Importance* (I) ratings are usually scored by multiplying the satisfaction and importance ratings (S × I), before summing responses concerning individual domains [Trauer & Mackinnon 2001]. As noted by Ferrans and Powers (1985), the rationale for multiplying satisfaction by importance ratings appears reasonable as individuals differ with regard to which dimensions predominate in importance. This implies that simple addition of satisfaction scores produces an inaccurate representation of QoL. Lehman (1982) has developed a widely-used QoL scale for mentally ill patients, in which he used results from previous studies. Other scales use methods such as interviews, focus groups, and expert opinions to generate relevant items to include [Juniper *et al.* 1996]. Again, some scales are constructed somewhat arbitrarily [Campbell *et al.* 1976].

Trauer & Mackinnon (2001) suggest that even though there is an intuitive appeal in weighting items by importance, this practice should be discontinued. Their views are shared by, e.g., Nunally (1967) who states *"For two reasons, it is usually not necessary to apply differential weights to the items on summative scales of attitudes (other than to reverse the scoring for negative statements). First, it is difficult to defend any particular method of weighting over the method of simply summing unweighting ratings. Second, and more to the point, weighted and unweighted summative scores usually correlate very* *highly."* To prove this, Likert (1932) undertook a study in which he compared unweighted scores with scores obtained by an elaborate method of weighting each item. The two sets of scores correlated with a coefficient of 0.99. On the other hand, Trauer & Mckin-non (2001) suggest, contrary to Nunally (1967), that using importance weights is not too difficult to defend: *"... the practice certainly appears to have considerable face validity. To be clear, importance may be a criterion in the initial selection of items for an instrument; we call into question the practice of multiplying included items by additionally obtained importance ratings."*

The importance of value placed on an item may also have explanatory value in its own right. For example, Oishi *et al.* (1999) find systematic differences in undergraduate students' level of well-being related to value orientation. Furthermore, Rejeski et al. (1998) find that the value that patients place on their physical function moderates their satisfaction ratings, that is, the greatest dissatisfaction is expressed by those who place a high value on their function despite conside-rable physical limitation. Thus even though there may be arguments for unweighted satisfaction ratings, this does not necessarily imply that all the items are of equal importance to the individual, nor that importance (or value) is not a valid and useful focus of study. Rather, problems arise in attempting to incorporate these ratings into the measurement of QoL using the multiplicative approach [Trauer & McKinnon 2001].

To sum up, it is suggested that satisfaction ratings already reflect a personal appraisal of the importance of the item to the individual, and that the multiplicative composite of satisfaction and importance has extremely undesirable measurement properties and also may be difficult to interpret. As concluded by Trauer & McKinnon (2001), there are no good reasons, beyond intuitive appeal, for multiplying satisfaction ratings by importance ratings, and consequently several compelling reasons for not doing so.

Results

Health status of the sample measured with SF-12

All respondents filled out the SF-12 profile questionnaire. The PF (Physical Function) scale includes 'moderate activities' and 'climbing several stairs'. Nearly 25 per cent of the respondents indicated that they were to some degree limited in performing moderate activities such as moving a table. Nearly 30 per cent were to some degree limited in climbing several flights of stairs. The RP (Role Function - Physical) includes the two attributes 'accomplished less' and 'limited in the kind of work' where over 40 per cent felt that they had somehow accomplished less than they would have liked during the last four weeks, and around 40 per cent were more or less limited in the kind of work or other activities during the last four weeks. Concerning the BP scale (Bodily Pain), nearly 40 per cent felt that some sort of pain interfered with their normal work during the past four weeks. The VT scale (Vitality) indicated that around 12 per cent of the respondents had a lot of energy during the last four weeks and around 7 per cent did not have any energy at any time. The SF scale (Social Function) indicated that around 20 per cent had some physical health or emotional problems during the last four weeks. The RE scale (Role Function-Emotional) contains the two attributes 'accomplished less' and ‘less careful than usual’. Over 40 per cent of the respondents felt that, to some degree, they had accomplished less than they would have liked during the last four weeks, and nearly 40 per cent felt that they, to some degree, had been less careful than usual regarding work during the last four weeks. The MH scale (Mental Health) encompasses the two attributes ‘calm and peaceful’ and ‘downhearted and blue’. Nearly 20 per cent of the respondents indicated that they had felt calm and peaceful all of the time and around 5 per cent had not felt so at any time. Over 50 per cent indicated that they, more or less, had felt downhearted and blue during the previous four weeks.

Subjective health status and sample characteristics

Table 1 illustrates the sample distribution of the general health (self-stated health) item in the SF-12 questionnaire according to gender and age. In total, around 85 per cent of the respondents rated their own general health to be at least ‘good’. None of the respondents rated their own general health as being ‘poor’. The differences between subjective health status as judged by males and females were insignificant at the 10 per cent level. The respondents’ age distribution in general was significant at the 1 per cent level, indicating that relatively elderly respondents assessed their subjective health as being worse than younger respondents.

Table 1. Distribution of respondents’ subjective assessment of their own health state by the SF-12 according to gender, age, and in total. Per cent in brackets.

	Excellent	Very good	Good	Fair	Poor	Total	*P*-value
Gender:							
Male	15 (6.9)	41 (18.9)	43 (19.8)	13 (6.0)	0 (0.0)	112 (51.6)	0.145*
Female	11 (5.1)	45 (20.7)	29 (13.4)	17 (7.8)	3 (1.4)	105 (48.4)	
Total	26 (12.0)	86 (39.6)	72 (33.2)	30 (13.8)	3 (1.4)	217 (100.0)	
Age (years):							
18 – 29	8 (3.7)	22 (10.2)	10 (4.6)	1 (0.5)	0 (0.0)	41 (19.0)	0.004*
30 – 39	4 (1.9)	16 (7.4)	8 (3.7)	5 (2.3)	1 (0.5)	34 (15.7)	
40 – 59	12 (5.6)	23 (10.6)	20 (9.3)	8 (3.7)	2 (0.9)	65 (30.1)	
≥ 60	2 (0.9)	24 (11.1)	34 (15.7)	16 (7.4)	0 (0.0)	76 (35.2)	
All	26 (12.0)	85 (39.4)	72 (33.3)	30 (13.9)	3 (1.4)	216 (100.0)	
In total	26 (11.6)	89 (39.6)	74 (32.9)	33 (14.7)	3 (1.3)	225 (100.0)	

*χ^2 test.

From Table 2 it can be seen that the mean score on the VAS scale for all respondents was around 80.0 with an interval ranging from 5.0 to 100.0. The mean VAS scores obtained for the two sub-samples were insignificant at the 10 per cent level. The mean VAS score for male respondents was higher than the corresponding score for females. This difference was significant at a 1 per cent level. Surprisingly, the mean VAS score did not differ significantly across age groups.

Table 2. VAS scores according to sample, gender and age. Per cent in brackets.

	VAS				
	Mean (SD)	Median	Min.	Max.	*P*-value
Sample:					
1 (n=99)	78.6 (20.8)	85.0	5.0	100.0	0.383*
2 (n=114)	80.6 (18.3)	90.0	15.0	100.0	

In total (n=213)	79.7 (19.5)	85.0	5.0	100.0	
Gender:					
Male (n=106)	82.2 (16.4)	87.5	15.0	100.0	0.003*
Female (n=102)	77.7 (22.1)	87.5	5.0	100.0	
Age (years):					
18-29 (n=40)	83.1 (13.8)	90.0	50.0	100.0	0.507**
30-39 (n=32)	77.2 (21.4)	85.0	30.0	100.0	
40-59 (n=62)	81.0 (20.0)	90.0	5.0	100.0	
≥ 60 (n=73)	78.4 (20.8)	85.0	15.0	100.0	

*Independent sample t-test.
**One-way ANOVA.

Table 3 illustrates socio-demographic data within the two samples and in total. The mean age of the respon-dents was 48.8 years. The difference according to mean age between the two samples was insignificant at the 10 per cent level. Around a quarter of the respondents lived alone, and fewer than five per cent of the respondents had five or more individuals in the household. The distribution of persons in the household in the two samples was insignificant at the 10 per cent level. Less than half of the respondents had a high school qualification and around 20 per cent had not been in school for more than maximum seven years. The distribution across the two samples was highly significant, indicating differences in the two samples due to the number of years the respondents had spent in both primary school and high school. Around 14 per cent of the respondents had a university degree, which is insignificant at the 10 per cent level for the two samples. Over 80 per cent of the respondents had a monthly income before tax below 30,000 DKK and around 1 per cent had a monthly income of 60,000 DKK or above. There were no significant differences in income between the two samples. Given that the lengths of the two questionnaires sent out to the two samples were different, it is not surprising that respondents in sample one took more time to complete the exercise, as this included the longer questionnaire. However, this difference in mean completion time was not statistically significant.

Table 3. Demographic variables by sample and in total.

	Sample 1 (n = 105)	**Sample 2 (n = 120)**	**All (n = 225)**	***P*-value**
Age, mean (n = 216)	49.3	48.3	48.8	0.765***
Persons in household (%):				0.732****
1 person	28.9	21.1	24.6	
2 persons	35.1	43.0	39.3	
3 persons	11.3	14.0	12.8	
4 persons	20.6	17.5	19.0	
5 persons	2.1	1.8	1.9	
≥ 6 persons	2.1	2.6	2.4	
Years in school (%):*				0.006****
≤ 7 years	13.4	22.9	18.6	
8-9 years	18.6	10.2	14.0	
10 years	13.4	27.1	20.9	
≥ 11 years	54.6	39.8	46.5	
University or college degree (%):				0.268****
Yes	16.8	11.5	13.9	
No	83.2	88.5	86.1	

Income before tax per month (%):				0.803****
≤ 19999 DKK**	60.0	55.2	57.6	
20000-29999 DKK	25.0	25.5	25.1	
30000-39999 DKK	8.0	11.4	9.8	
40000-49999 DKK	5.0	6.1	5.6	
50000-59999 DKK	1.0	0.0	0.5	
≥ 60000 DKK	1.0	1.8	1.4	
Time to fill out questionnaire (min.):				
Mean (SD)	30.0 (15.8)	27.6 (12.3)	28.7 (14.1)	0.576***
Median	25.0	25.0	25.0	
Minimum	10.0	5.0	5.0	
Maximum	90.0	70.0	90.0	

*Including primary and high school.

**1 DKK equals 0.1347 Euro.

***Independent sample *t*-test.

****χ^2 test.

Results from the MAUT exercise

Only 225 respondents returned the questionnaire, resulting in a return rate of only 11 per cent, which was very low considering the usual return rate for postal-based questionnaires in Denmark. Nevertheless, the valid answers were judged usable for the calculations of weights for models (1) and (2). In pair-wise comparisons, most of the mean Kj weights from the top levels differed significantly from those from the bottom levels (cf. Table 4).

Table 4. The mean K_j weights for the dimensions from the two samples.

Dimension*	**K_{j1} (n = 105)**	**K_{jb2} (n =120)**
1	0.0914	0.0922
2	0.0831	0.1000
3	0.0972	0.1020
4	0.0996	0.0960
5	0.0945	0.0963
6	0.0996	0.0946
7	0.0964	0.0839
8	0.0850	0.0897
9	0.0782	0.0965
10	0.0788	0.0774
11	0.0962	0.0714
Σ	1.0000	1.0000

*Only eleven dimensions were valued since the dimension ’Assessment of own health” could not be fitted into the models.

Table 5. Mean scores (SD) and 95% confidence intervals across levels for all dimensions.

Dimension	Level 1	Level 2	Level 3	Level 4	Level 5
1	1=1.0000	2=0.6052 (0.255) [0.5540; 0.6322]	3=0.2088 (0.247) [0.1690; 0.2390]	4=irrelevant	4=irrelevant
2	1=1.0000	2=0.6245 (0.227) [0.5838; 0.6542]	3=0.2610 (0.254) [0.2181; 0.2926]	5=irrelevant	5=irrelevant
3	1=1.0000	2=0.6979 (0.225) [0.6690; 0.7354]	3=0.4744 (0.220) [0.4443; 0.5113]	4=0.3118 (0.214) [0.2829; 0.3492]	5=0.1755 (0.221) [0.1500; 0.2199]
4	1=1.0000	2=0.6816 (0.232) [0.6378; 0.7114]	3=0.4559 (0.213) [0.4184; 0.4839]	4=0.2990 (0.218) [0.2630; 0.3304]	5=0.1519(0.210) [0.1227; 0.1894]
5	1=1.0000	2=0.6732 (0.256) [0.6338; 0.7115]	3=0.4795 (0.237) [0.4458; 0.5180]	4=0.3222 (0.242) [0.2816; 0.3540]	5=0.1802 (0.240) [0.1473; 0.2225]
6	1=1.0000	2=0.6546 (0.259) [0.6125; 0.6922]	3=0.4550 (0.229) [0.4209; 0.4915]	4=0.2934 (0.225) [0.2619; 0.3327]	5=0.1579 (0.233) [0.1285; 0.2024]
7	1=1.0000	2=0.6660 (0.247) [0.6249; 0.7014]	3=0.4803 (0.236) [0.4482; 0.5218]	4=0.3188 (0.238) [0.2821; 0.3558]	5=0.1670 (0.236) [0.1321; 0.2066]
8	1=1.0000	2=0.8159 (0.196) [0.7727; 0.8367]	3=0.6478 (0.230) [0.6091; 0.6804]	4=0.4718 (0.244) [0.3878; 0.4627]	5=0.1972 (0.257) [0.1625; 0.2414]
9	1=1.0000	2=0.8124 (0.188) [0.7717; 0.8329]	3=0.6642 (0.204) [0.6313; 0.6961]	4=0.4258 (0.220) [0.4002; 0.4670]	5=0.1919 (0.250) [0.1647; 0.2420]
10	1=1.0000	2=0.7036 (0.261) [0.6625; 0.7423]	3=0.4861 (0.234) [0.4591; 0.5313]	4=0.2938 (0.235) [0.2698; 0.3434]	5=0.1463 (0.240) [0.1189; 0.1953]
11	1=1.0000	2=0.6712 (0.273) [0.6330; 0.7144]	3=0.4993 (0.257) [0.4676; 0.5457]	4=0.3012 (0.260) [0.2645; 0.3433]	5=0.1555 (0.269) [0.1186; 0.2020]

Note: NA (Not Applicable) refers to the fact that dimensions 1 and 2 comprise only three levels.

Table 5 shows the 'social' weights for the levels across all dimensions. All levels were estimated by group means, however, there was a fairly high variation for both upper levels (best) and lower levels (worse). While the relative values of the various levels of the dimensions were the same in both models, the relative importance weights (K_j) were different. This is because model (2) took into account that relative importance may change as a function of levels. Nevertheless, when the equations for models 1 and 2 were used on the respondents' profiles (results from the SF-12 questionnaire), the differences were minor (cf. Table 6).

Table 6. Descriptive statistics of the values for models (1), (2) and self-reported health scores (using VAS).

Variable	Mean (SD)	Median	Min	Max	N	% < 0.32
V_{HM1}	0.825 (0.154)	0.868	0.320	1.000	209	0.0
V_{HM2}	0.826 (0.154)	0.870	0.321	1.000	209	0.0
Self-reported	0.797 (0.195)	0.850	0.050	1.000	213	2.1

As can be seen in Table 6, the means for both models (1) and (2) were higher than the means from the self-reported health states using the VAS method. The worst score, according to models 1 and 2, was 0.32. The worst score according to the VAS was 0.05. Only around 2 per cent of the VAS scores were below the minimum scores for models (1) and (2).

Correlation coefficients for models (1), (2) and VAS results

Since there is no existing preference index for SF-12, a test for construct validity cannot be performed, and hence one cannot get an indication of the validity of the results stemming from the two models. Instead the results from the VAS exercise were used to get an impression of how close the mean values obtained by the two models corresponded with the results from the VAS exercise, i.e. the VAS mean is effectively treated as the gold standard. Using Pearson and Spearman correlation coefficients it appears that, as illustrated in Tables 7 and 8, the correlation coefficients between A) the two models, B) model (1) and results from the VAS exercise, and C) model (2) and results from the VAS exercise, were all significant at the 1 per cent level.

Table 7. Pearson correlation coefficients between results from model (1), model (2), and VAS.

	Model (1)	Model (2)	VAS
Model (1)	1.000		
Model (2)	1.000*	1.000	
VAS	0.788*	0.788*	1.000

*(p < 0.01).

Table 8. Spearman correlation coefficients between results from model (1), model (2), and VAS.

	Model (1)	Model (2)	VAS
Model (1)	1.000		
Model (2)	1.000*	1.000	
VAS	0.683*	0.682*	1.000

*(p < 0.01).

Discussion

A limitation of both generic and disease-specific quality-of-life instruments is that they are mainly descriptive, i.e., they describe health status on a number of different dimensions. However, for evaluative purposes it would be useful to measure overall health-related quality of life along a cardinal scale. SF-12 is no exception, and it was the aim of this study to provide health-state utilities and, subse-quently, an algorithm that could be applied in future studies to generate a single index score when using SF-12 for health status measurement.

This study should only be considered an experiment, and consequently we limited our models to encompass two additive models. Other model forms, e.g. multiplicative or multi-linear, may perform better. However, as shown in a similar study using the 15D instrument, the two additive models behaved very well in the case of applying MAUT for valuations of the 15D, which led us to limit ourselves to those two models in this first preliminary study [Wittrup-Jensen & Pedersen 2001].

Due to the structure of the SF-12 we had to exclude the item 'General Health (GH)' from the valuation task. This conclusion was also reached by Lundberg *et al.* (1999), who also excluded this item in their valuation study.

In estimating these preliminary results we interpreted the SF-12 as consisting of eleven unique and inde-pendent questions, i.e. dimensions. However, the SF-12 consists of only six dimensions, where in some instances two questions represent one dimension. Hence it could be argued that the mean K_j weights ought to be expressed for six dimensions only and not, as presented here, for eleven dimensions. In other words, some sort of weighting process should be incorporated. In practice this could be accomplished by using either *Rasch models* or *factor analysis*. Conse-quently, this issue should be investigated more deeply in future studies eliciting health-state utilities for the SF-12.

Both models (1) and (2) resulted in an algorithm which makes it possible to apply this algorithm in future studies and elicit a single index score that can be applied in economic valuations. The algorithms are based on the MAUT technique, which has been used in eliciting health-state utilities for both the 15D and the Health Utility Index (HUI) [Sintonen 1994; Torrance *et al.* 1995]. Lundberg *et al.* (1999) suggest that the MAUT method could be used as an alternative to using regression analysis, and that which approach is to be preferred for eliciting health-state utilities depends on which approach can most closely predict the results of directly measuring health-state utilities. However, the MAUT method has some drawbacks, especially in the sense that we make use of category scaling and magnitude estimation. Neither of these methods explicitly incorporates a trade-off, which may disqualify them as proper scaling techniques for eliciting individual utilities. According to Richardson (1994), both methods make it difficult in placing a meaning on the units obtained from these scales, and the difficulty in judging their suitability as a basis for allocating resources remains. We agree that it is questionable whether these scales yield interval properties, which are required to permit the summation of utilities (QALYs). Nevertheless, the methods have some advantages, especially in cases where the health status instrument is relatively comprehensive as is the case with the SF-12 and 15D, where other scaling techniques are inappropriate. Actually, this inappropriateness is not due to theoretical limitations, but merely due to the fact that it is nearly impossible to obtain direct valuations from so many health states, which makes it hard to predict - at least with a reasonable fit - the remaining health states. Evidently this limitation is not about the appropriateness of the scaling techniques, but about budget constraints.

There have been some efforts to elicit health-state utilities for the SF-12. The first method is in progress and not fully developed. This is the method proposed by Brazier and Roberts (2001), which is based on the Standard Gamble (SG) scaling technique. The idea is to estimate a sub-sample and apply regression analysis in order to predict the remaining possible health states. At first glance this is the same procedure used in Dolan (1997) and Wittrup-Jensen *et al.* (2001) when they estimated health-state utilities for the EuroQol (EQ-5D). However, there is a major difference, since the SF-12 encompasses around 7,500 possible health states and the EQ-5D only 243 health states. Brazier and Roberts (2001) elicited direct valuations of 249 SF-12 health states, where each health state was valued - on average - by six different respondents. Based on these direct valuations, the remaining non-direct health states were predicted by applying regression analysis. In other words, they used 3.3 per cent of the health states to predict the remaining 96.7 per cent. Although they did take account of obtaining direct valuations of a variety of different health states covering the whole spectrum of the SF-12, the robustness of the prediction cannot be very high, *ceteris paribus*. Nevertheless, the fit of the regression analysis yielded an R^2 value of 0.426, which - everything taken into consideration - is modest and ought to be higher. We believe that the method adopted by Brazier and Roberts (2001) has its advantages and that it all comes down to whether one is for or against the SG as a proper scaling technique in eliciting utilities which can be entered in economic evaluations. However, we would like to stress that eliciting direct valuations from more than these 3.3 per cent of health states, all other things being equal, would lead to an improvement of both the fit and consequently the robustness of the analysis. Finally, each health state was only valued - on average - by six respondents. There is no reason - apart from having a limited budget - why this number should not be raised to 12-14 respondents, which would improve the robustness of the analysis.

The second study has been undertaken by Lundberg *et al.* (1999), who used the RS and TTO to elicit health state utilities. They concluded that their regression models explained about 50 per cent of the variance in the RS responses and about 25 per cent in the TTO responses. The RS has some known limitations in eliciting individual utilities due to a weaker theoretical foundation than the TTO and the SG methods [Bleichtrodt and Johannesson 1997]. The R^2 value for the TTO in the study performed by Lundberg et al. (1999) is much lower than the R^2 value obtained by using the SG me-thod in the study undertaken by Brazier and Roberts (2001). There may be many reasons for this. One reason is that Lundberg et al. used a postal-based questionnaire. In our experience the TTO method is too difficult and cumbersome to be applied in a postal-based study. Instead it performs very well when applied in an interview-based study. This may to some degree explain the low R^2 value. Lundberg *et al.* (1999) recommend that future studies should use interview-based scenarios instead of postal-based. The study by Lundberg et al. evidently suffers from the same limitations as the study by Brazier and Roberts, since they also had to estimate a fairly high number of predicted valuations based on relatively few direct valuations.

The mean K_j weights for the eleven dimensions for both models (1) and (2) were around 0.10 with a minimum of 0.0782 and maximum of 0.0996 for model (1), and 0.0714 and 0.1020 for model (2). The mean scores for the levels varied across dimensions, which indicated that respondents put more emphasis on some levels in a dimension compared to other levels in other dimensions. This was also the case in the previous chapter, where we estimated weights and mean scores for the 15D.

A major limitation to our study is the lack of testing for validity. Due to the fact that we did not use regression analysis, we were unable to estimate an explicit indicator for the fit, e.g. a R^2 value. However, another way to test for validity is to look at construct validity. This approach has its drawbacks as well, since one necessarily has to test one's results against a *gold standard*, which in our case does not exist. The line of reasoning in this case is that one simply declares an alternative valuation method to be the gold standard and then explicitly compares the results from one's study with the results from the alternative study. Given that the algorithm developed by Brazier and Roberts (2001) is currently unavailable we cannot test our results against results estimated by their algorithm. Another way to test for validity is to use correlation analysis. We estimated correlation coefficients by comparing results from the two models and results from the VAS exercise. Not surprisingly, all correlation coefficients proved to be statistically significant. Nevertheless, we would like to point out that a significant correlation coefficient does not necessarily have to be interpreted as a high degree of validity, and that the results from the correlation analysis should be interpreted with a grain of salt. In this particular case, first due to the appropriateness of the VAS scale in itself, and second the correlation analysis in itself is dubious.

Preliminary results show that it is possible to use MAUT for the SF-12. It appears that the additive model provides a reasonable approximation to how people value multi-attribute health states. However, we urge that future research investigate other model specifications e.g. multiplicative, in addition. Both models generated only positive scores, i.e. all possible health states were thus regarded as preferable to death, and our design did not include the states 'dead' or 'unconscious'. Of course it is possible to scale the scores so that negative ones are also allowed. However, whether these should be allowed is a matter of perspective. As in Torrance *et al.* (1992) and Sintonen (1994) we also restricted the scores to the positive area.

In conclusion, these findings are very preliminary and much more research applying MAUT to the SF-12 is needed. However, we feel that the MAUT method is a valid alternative to existing preference elicitation strategies for SF-12 and that our proposed valuation technique can be followed in estimating an official preference-index for the SF-12.

References

1. Bleichrodt H & Johannesson M. An experimental test of a theoretical foundation for rating scales valuations. *Medical Decision Making* 1997; 17: 208-16.
2. Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 Health Survey. *Journal of Clinical Epidemiology* 1998; 51(11): 1115-1128.
3. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics* 2002; 21: 271-292.
4. Brazier J & Roberts J. *Estimation of a preference-based index measure of health for the SF-12 and comparison to the SF-36 preference-based index.* Poster presentation at the iHEA meeting in York, UK. 2001.
5. Campbell A, Converse PE, Rodgers WL. *The quality of American life.* Russell Sage Foundation: New York. 1976.
6. Devlin NJ, Hansen P, Kind P, Williams A. *The health state preferences and logical inconsistencies of New Zealanders: A tale of two tariffs.* York Centre for Health Economics UK and University of Otago New Zealand. Discussion Paper No 180. 2000.
7. Diener E. Subjective well-being. *Psychological Bullentin* 1984; 95: 542-575.
8. Dolan P. Modelling valuations for EuroQol health states. *Medical Care* 1997; 35(11): 1095-1108.
9. EuroQol Group The. EuroQol: A new facility for the measurement of health-related quality of life. *Health Policy* 1990; 16: 199-208.
10. Feeny D, Furlong W, Boyle M et al. Multi-attribute health status classification systems: Health Utilities Index. *PharmacoEconomics* 1995; 6: 490-502.
11. Ferrans CE & Powers MJ. Quality of life index: Development and psychometric properties. *Advanced Nursing Science* 1985; 8(1): 15-24.

12. Fryback DG, Lawrence WF, Martin PA, Klein R, Klein BE. Predicting quality of well-being scores from the SF-36: Results from the Beaver Dam health outcomes study. *Medical Decision Making* 1997; 17: 1-9.

13. Gold M, Franks P, Erickson P. Assessing the health of the nation. The predictive validity of a preference-based measure and self-rated health. *Medical Care* 1996; 34(2): 163-177.

14. Green SB, Salkind NJ, Akey TM. *Using SPSS for Windows*. Printice Hall: New Jersey. 1997.

15. Johnson JA & Coons SJ. Comparison of the EQ-5D and SF-12 in an adult US sample. *Quality of Life Research* 1998; 7: 155-166.

16. Juniper EF, Guyatt GH, Jaeschke R. How to develop and validate a new health-related quality of life instrument. In Spilker B (eds) *Quality of life and pharmacoeconomics in clinical trials*. Lipppincott-Raven: Philadelphia. 1996.

17. Lehman AF, Ward NC, Linn LS. Chronic mental patients: The quality of life issue. *American Journal of Phychiatry* 1982; 139(10): 1271-1276.

18. Lundberg L, Johannesson M, Isacson DG, Borgquist L. The relationship between health-state utilities and the SF-12 in a general population. *Medical Decision Making* 1999; 19: 128-140.

19. McHorney CA, Ware JE, Raczek AE. The MOS 36-item short-form health status survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Medical Care 1993; 31: 247-.

20. Nunally JC. *Psychometric theory.* McGraw-Hill: New York. 1967.

21. Nunally JC & Bernstein IR. *Psychometric theory*. McGraw-Hill: New York. 1994.

22. Oishhi S, Deiner E, Suh E, Lucas RE. Value as a moderator in subjective well-being. *J Personal* 1999; 67(1): 157-184.

23. Pliskin JS, Shepard DS, Weinstein MC. Utility functions for life years and health status. *Operations Research* 1980; 28(1): 206-224.

24. Rejeski WJ, Martin KA, Miller ME, James MK, Rapp WHJ, Messier SP. Validation of the PASE in older adults with knee pain and physical disability. *Medical Science & Sports Exer* 1999; 31(5): 627-633.

25. Richardson J. Cost utility analysis: What should be measured? *Social Science and Medicine* 1994; 39(1): 7-21.

26. Shmueli A. Subjective health status and health values in the general population. *Medical Decision Making* 1999; 19: 122-127.

27. Sintonen H. An approach to measuring and valuing health states. *Social Science and Medicine* 1981; 15: 55-65.

28. 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, Melbour-ne. 1994.

29. Torrance GW, Boyle MH, Horwood SP. Application of multi-attribute theory to measure social preferences for health states. *Operations Research* 1982; 30: 1043-1069.

30. 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.

31. Torrance GW, Furlong W, Feeny D, Boyle M. Multi-attribute preference functions. Health utilities index. *PharmacoEconomics* 1995; 6: 503-520.

32. Trauer T & Mackinnon A. Why are we weighting? The role of importance ratings in quality of life measurement. *Quality of Life Research* 2001; 10: 579-585.

33. Ware JE, Snow KK, Kosinski M et al. *SF-36 health survey: Manual and interpretation guide*. Boston, MA: The Health Institute, New England Medical Center. 1993.

34. Ware JE, Kosinski M, Keller SD. *SF-36 physical and mental health summary scales: A user's manual*. Boston, MA: The Health Institute, New England Medical Center. 1994.

35. Ware JE, Keller SD, Gandek B et al. Evaluating translations of health status surveys: Lessons from the IQOLA project. International Journal of Health Technology Assessment 1995A; 11: 525-.

36. Ware JE, Kosinski M, Bayliss MS et al. Comparison of methods for the scoring and statistical analysis of SF-36 health profiles and summary measures: Results from the medical outcomes study. Medical Care 1995B; 33(4): 264.

37. Ware JE jr., Kosinski M, Keller SD. A 12-item short-form health survey. Construction of scales and preliminary tests of reliability and validity. Medical Care 1996; 34(3): 220-233.

38. Ware J & Gandek B. Overview of the SF-36 health survey and the international quality of life assessment (IQOLA) project. *Journal of Clinical Epidemiology* 1998; 51(11): 903-912.

39. Wittrup-Jensen KU & Pedersen KM. *Modelling weights for 15D*. Paper presented at the 22nd Nordic Study Group Meeting in Odense, 22 – 25 August 2001.

40. Wittrup-Jensen KU, Lauridsen JT, Gudex C, Brooks R, Pedersen KM. *Danish EuroQol tariffs estimated by the Visual Analogue Scale (VAS) and the Time Trade-Off (TTO).* Paper presented at the 6th EuroQol Group Meeting in Copenhagen, 6th/7th September 2001.