CENTRE FOR HEALTH PROGRAM EVALUATION

WORKING PAPER 42

The 15-D Measure of Health Related Quality of Life. II Feasibility, Reliability and Validity of its Valuation System

Harri Sintonen
Professor,
Department of Health Policy and Management, University of Kuopio, Finalnad,
and
Visiting Fellow, National Centre for Health Program Evaluation

February, 1995
ISSN 1038-9547
ISBN 1 875677 37 2

CENTRE PROFILE

The Centre for Health Program Evaluation (CHPE) is a research and teaching organisation established in 1990 to:

- undertake academic and applied research into health programs, health systems and current policy issues;
- develop appropriate evaluation methodologies; and
- promote the teaching of health economics and health program evaluation, in order to increase the supply of trained specialists and to improve the level of understanding in the health community.

The Centre comprises two independent research units, the Health Economics Unit (HEU) which is part of the Faculty of Business and Economics at Monash University, and the Program Evaluation Unit (PEU) which is part of the Department of General Practice and Public Health at The University of Melbourne. The two units undertake their own individual work programs as well as collaborative research and teaching activities.

PUBLICATIONS

The views expressed in Centre publications are those of the author(s) and do not necessarily reflect the views of the Centre or its sponsors. Readers of publications are encouraged to contact the author(s) with comments, criticisms and suggestions.

A list of the Centre's papers is provided inside the back cover. Further information and copies of the papers may be obtained by contacting:

The Co-ordinator
Centre for Health Program Evaluation
PO Box 477
West Heidelberg Vic 3081, Australia
Telephone + 61 3 9496 4433/4434 **Facsimile** + 61 3 9496 4424
E-mail CHPE@BusEco.monash.edu.au

ACKNOWLEDGMENTS

The Health Economics Unit of the CHPE receives core funding from the National Health and Medical Research Council and Monash University.

The Program Evaluation Unit of the CHPE is supported by The University of Melbourne.

Both units obtain supplementary funding through national competitive grants and contract research.

The research described in this paper is made possible through the support of these bodies.

AUTHOR ACKNOWLEDGMENTS

The financial support of the Academy of Finland for this study is gratefully acknowledged. My special thanks are to Markku Pekurinen for his invaluable contribution to the development of 15D in many ways. I am also indebted to the staff of the Department of Public Health, University of Helsinki for assistance in carrying out the valuation survey. My special thanks are to the National Centre for Health Program Evaluation for excellent working environment while preparing this paper and to Professor Jeff Richardson for valuable comments.

ABSTRACT

The 15D is a generic, 15-dimensional, standardised, self-administered measure of health-related quality of life (HRQOL), that can be used as a profile and single index score measure. This paper introduces alternative valuation models based on the multi-attribute utility theory for generating the single index scores, and examines their feasibility, reliability and validity by using several data sets and methods. Valuations were elicited from several representative samples of Finnish adult population by using rating scales in self-administered questionnaires and postal survey. The approach proved to be feasible and produced valuations with a high reliability. There is solid convergent evidence of construct validity for the index scores generated by additive 2-stage or 3-stage valuation models. Tests for reflective equilibrium provided strong evidence that these scores exhibit a more plausible trade-off between length and quality of life than some other single index score measures (EuroQol, QWB, McMaster), and are thus more valid for QALY calculations in cost-utility analysis. The values are consistent and quite likely widely generalisable and usable at least in Western-type societies. The 15D is thus probably the most sensitive and comprehensive HRQOL measure presently available that combines the advantages of a profile and single index score measure with a high level of reliability and validity.

The 15-D Measure of Health Related Quality of Life. II Feasibility, Reliability and Validity of its Valuation System

1 Introduction

In an earlier paper the health state descriptive systems of two versions of the 15D measure of health-related quality of life were introduced and their properties as a profile measure examined (Sintonen 1994). It was found that the 15D performs very well in that capacity.

Profiles describe what one's health-related quality of life (HRQOL) is like. They show on which dimensions differences between individuals and groups occur cross-sectionally and what changes on them take place over time. However, if one profile does not dominate another, that is, the score is better on each dimension, but the profiles cross (one individual or group is better-off on some dimensions, but worse-off on others), it is impossible to say, which individual or group is better-off in HRQOL overall cross-sectionally, or if inferences are made over time, whether, as a whole, the HRQOL has improved or deteriorated.

With profiles it may thus not be even possible to make ordinal comparisons of goodness or badness, let alone cardinal ones, that is, to say how much better or worse a profile is than another. Relative goodness or badness is a value judgment. Therefore a value component is needed to aggregate the measurements on the dimensions into a single index score. The value component should reflect a quantitative `social' value judgment of the overall goodness or badness (quality weights) of various profiles (health states). `Social' is usually taken to imply that the values reflect those of an appropriate group of people.

The usefulness of instruments having both a profile and single index score feature has been described as follows: `Quality of life research might profit from such instruments given that they have been developed according to a defined conceptual basis, to rigorous quality standards and to easy applicability. Their value, for example in clinical trials, lies in providing both detailed information on quality of life components affected differently by the treatment arms and a single measure of overall effect. This single measure would be optimal if derived from well thought-out weighting operations and if usable as an unequivocal scale with defined segments. Depending on the study objectives, either the profile part or the index part of the instrument could be chosen for hypothesis testing' (Bullinger 1993, pp. 218-219).

There are several methods for eliciting values for health states from individuals. The most frequently used techniques are rating scale (RC), magnitude estimation (ME), standard gamble (SG), time trade-off (TTO) and equivalence technique or person trade-off (PTO). The methods have been described in detail elsewhere (Torrance 1986, Froberg and Kane 1989, Nord 1994). They seem to produce different values for the same states (Patrick et al. 1973, Torrance 1976, Read et al. 1984, Gudex et al. 1993). There are several reasons for this variability, all of which are far from fully understood (Nord 1992), but the major problem is that none of the methods can be

regarded as a gold standard. So what method to use?

In connection with the 15D health state descriptive system, there is no point in discussing that issue generally. The above methods are holistic and direct in nature in the sense that the health states to be valued are described to the subjects as a whole. The 15D descriptive system defines an enormous number of mutually exclusive 15-dimensional health states. It is plainly obvious that in this case none of the methods, at least as applied usually, is feasible: the complexity (cognitive overload) and number of health state descriptions make it impossible to fill the valuation space adequately.

The purpose of this paper is to describe a feasible approach to producing alternative value components for the 15D and to evaluate their properties theoretically and empirically. The value components are described in section 2. Feasibility is clearly a necessary, but not a sufficient condition for a valuation method: there are also other criteria that should be met adequately. These criteria and the materials and methods used to elicit the value components for the 15D.2 and to evaluate them are introduced in section 3. The empirical results are presented in section 4 and discussed in section 5.

2 Alternative Value Components for the 15D

If we want a standardised and sensitive single index score measure, the multi-attribute utility (MAU) method (Keeney and Raiffa 1976) may provide a feasible solution to the valuation problem of a large number of health states. In the first version of 15D (it was 12D then) the method was applied by using the following two-stage additive valuation model (Sintonen 1981b):

$$v_{HM1} = \sum_j l_j[w_j(x_j)], \qquad (1)$$

where v_{HM1} = the `social' value of health state H as produced by model (1), l_j = a positive constant for the jth dimension (j=1, 2,...,m), representing its relative importance ($\Sigma l_j = 1$), and $w_j(x_j)$ = a numerical function of the jth dimension, representing the relative value of the various levels of the dimension (top level = 1, being dead = 0).

This model is used also here as the simplest alternative. It assumes that the dimensions are valuewise additive independent (Keeney and Raiffa 1976) and the importance weights apply over the whole range of levels, that is, the difference in the relative importance between any two dimensions remains constant. To what extent these assumptions hold will be tested later. Yet a priori it would appear more plausible that the relative importance may change as a function of levels. A model that allows for that could be of the form:

$$v_{HM2} = \sum_j [l_j(x_j)][w_j(x_j)], \qquad (2)$$

where $l_j(x_j)$ = a set of positive constants for the jth dimension, representing the relative importance of the dimension at its various levels ($\Sigma l_j = 1$ at any level), and $w_j(x_j)$ = a numerical function on the jth dimension, representing the relative value of the various levels of the dimension (top level = 1, being dead = 0). Also (2) assumes additive independence.

Values produced by models (1) or (2) have been used in all 15D applications so far. However, still

other model formulations are possible. Torrance et al. (1982) used a multiplicative (dis)utility model of the form:

$$\bar{U}_{HM3} = 1 - s_{M3}\bar{u}_{HM3}, \quad (3)$$

where u_{HM3} = the `social' utility of health state H as produced by model (3), s = a scaling factor, $\_{HM3}$ = the `social' disutility of health state H defined as

$$u_{HM3} = (1/c)\{\prod_j [1+c\cdot c_j[u_j(x_j)]] - 1\} \quad (3a)$$

and $u_j(x_j)$ = a numerical function on the jth dimension, representing the relative utility of the various levels of the dimension (top level = 1, bottom level = 0). The c_j values resemble the importance weights in (1) and (2) with the distinction that they are not scaled to sum up to 1, but their sum has a relation to the interaction parameter c as follows:

- if $\Sigma c_j > 1$, then $-1 < c < 0$ (dimensions are `substitutes'), (3b)
- if $\Sigma c_j = 1$, then $c = 0$, and the additive model holds, and (3c)
- if $\Sigma c_j < 1$, then $c > 0$ (dimensions are `complements'). (3d)

This `utility' model differs from the `value' models (1) and (2) also in that the underlying `value' model was nonlinearly converted to a scale that presumably would have been obtained by using the SG method in valuation. Torrance et al. (1982) used a conversion factor of $u = 1-(1-v)^{1.6}$ and allowed also negative health state utilities. With these features model (3) is applied to the 15D descriptive system and tested here.

Another alternative to be tested is (3) without the `utility' conversion of the form:

$$v_{HM4} = 1 - z_{M4}v_{HM4}, \quad (4)$$

where v_{HM4} = the `social' value of health state H as produced by model (4), z = a scaling factor, $v_{HM4}$ = the `social' disvalue of health state H defined as

$$v_{HM4} = (1/d)\{\prod_j [1+d\cdot d_j[w_j(x_j)]] - 1\} \quad (4a)$$

and $w_j(x_j)$ = a numerical function on the jth dimension, representing the relative value of the various levels of the dimension (top level = 1, bottom level = 0).

Recently, Torrance et al. (1992) modified (3) in two important respects. They used the SG directly to derive utilities and obtained an estimated utility conversion factor of $u=1-(1-v)^{2.29}$. Second, they did not allow negative utilities. A utility model for 15D based on these features is referred here to as model (5) and the corresponding value model as model (6).

Models (1)-(6) produce a `social' value or `utility' to all health states defined by the 15D descriptive system on a 0-1 scale (being dead=0, no problems on any of the dimensions=1).

3 Materials and Methods

The valuation tasks

For model (1) essentially the same methodology and format was used as described in Sintonen (1981b). On one page, the dimensions were described by their top levels (level 1) adjacent to a vertical continuous 0-100 ratio scale. 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 valuation task (task 1) was introduced to the subjects (i= 1, 2,...,n) as follows:

`Below is a list of attributes related to health. People may, however, have different views about how important they are from the viewpoint of health. Here we are interested in your opinion.

Please assess first, which of the attributes below is in your opinion ***the most important one*** *from the viewpoint of health, that is, the one that you would like to give up last. Then draw a line from the box (•) following it to 100 on the adjacent scale. Then assess* ***the importance of all other attributes in relation to this most important attribute****. If, for example, an attribute is in your opinion half (1/2 or 50%) as important as the most important one, draw a line from the box following it to 50 on the scale. If an attribute is in your opinion not at all important, draw a line from its box to 0. For Clarity, please write in each box the number, at which the line is aimed (eg). In the assessm* 50 *u can use all numbers between 0 and 100 as you see fit. The lines can cross.'*

The ratio scale nature of the valuation task was further emphasised by placing to the right hand side over the range of the scale nine arrows with text explaining, how the number pointed by the arrow should be interpreted. For example, the arrow pointing to 90 read: 9/10 as important as the most important attribute (90% of the importance of the most important attribute). The values obtained were divided by 100 to bring them to a 0-1 scale and then transformed to satisfy $\Sigma_j l_{ij} = 1$. `Social' importance weights (l_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. The five levels plus the states `being unconscious' and `being dead' as the two bottom ones, were described adjacent to a 0-100 ratio scale of desirability. The instructions were the following (task 2):

`On this and the next 7 pages different health states are presented. You should assess, how desirable the states presented on each page are in relation to each other. Please read these instructions with care, as they apply on each of the next 7 pages.

At a time, always think of the health states, always think only of the health states that are presented on that page (above or under the line). From the box following the topmost state, which obviously is in everybody's opinion the most desirable one, a line has been drawn to 100 on the adjacent scale. You should now ***assess the desirability of all other states in relation to this most desirable one.*** *If, for example, a state is in your opinion half (1/2 or 50 %) as desirable as the most desirable state, draw a line from the box following it to 50 on the scale. From the box of the least desirable state, draw a line to 0. For clarity, please write in each box the number, at which the line is aimed (eg). In the assessment you* 50 *se all numbers between 0 and 100 as you see fit.'*

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) and $w_{ij}(x_{jD}) = 0$ (dead). `Social' level values $w_j(x_j)$ were calculated by averaging them over the sample.

Three further versions of task 2 with an identical format were used. Task 3 was similar to task 2,

but the duration of the states was defined to be 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 4 the duration was one month. In task 5 the duration was again one year and the states `unconscious' and `dead' were not included as the two bottom levels.

For model (2) weights for the bottom level of each dimension (level 5 for 15D.2) were elicited with a format resembling that of the EuroQol (EuroQol Group 1990). The instructions were (task 6):

`Below is a list of some health states, where people can be in. People may have different views about how good or bad these states are. Here we are interested in your views.

Please draw a line from the box following each state (•) to the point on the adjacent scale, which shows, how good or bad that state is your opinion is ***in relation to the best imaginable and worst imaginable health state.*** *The best imaginable health state is marked by 100 and the worst imaginable one by 0 on the scale. For clarity, please write in each box the number, at which the line is aimed (eg). In* 50 *sessment you can use all numbers between 0 and 100 as you see fit. The lines can cross.'*

Here again the duration of the states was unspecified. The values obtained were divided by 100 and transformed to satisfy $\Sigma_j l_{jb} = 1$ (b refers to the bottom level of j). `Social' weights ($l_{jb}$) were formed by averaging the individual weights over the sample. 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 2. In another version of this task, the duration was defined to be one year (task 7).

In models (3)-(6) the group solution suggested by Torrance et al. (1982) was applied. The $w_j(x_j)$ values on a 0-1 scale (top level = 1, bottom level (level 5) = 0) were obtained from tasks 2 and 5 (see results), and converted to `utilities' by using $u=1-(1-v)^{1.6}$ for model (3) and $u=1-(1-v)^{2.29}$ for model (5). The values for level 5 of each dimension from tasks 6 and 7 were used to derive the `utilities' for these corner states and the fifteen c_j values for models (3) and (5) (and the d_j values for models (4) and (6)). For models (3) and (4) the value for the combination of the worst level (level 5) from each dimension in relation to the states of `being unconscious' and `being dead' was derived by using a EuroQol-type format as follows: (task 8):

`Even if you may find this task difficult, we would like you to compare the states of `being unconscious' and `being dead' with the state, that is the combination of the bottom levels (level 5) of all the previous 15 questions (described in the big box). Please imagine that the state `unconscious' and the combination state last for one year. What happens after that is not known and should not be taken into account.

Please draw a line from the box following each state (•) to the point on the adjacent scale, which shows, how good or bad that state in your opinion is ***in relation to the best imaginable and worst imaginable health state.*** *The best imaginable health state is marked by 100 and the worst imaginable one by 0 on the scale. The task may be easier, if you first decide, which of the tree alternatives is the worst one. Then assign to it a value on the scale on the basis of how close to the worst imaginable state you think it is.'*

The values were transformed to a 0-1 scale with the best imaginable health state (presumably the combination of the best levels) = 1 and `being dead' = 0. The value of the worst combination was then converted to a `utility' for model (3). The calculations required for (3)-(6) are described in detail in Torrance et al. (1982) and (1992).

Finally, a scale similar to that described in task 6 was used by the subjects to assess their own overall health status in the day they filled in the questionnaire (task 9).

The samples

All valuation tasks were carried out with self-administered postal questionnaires with one reminder and a new questionnaire sent about two weeks after the original mailing. Five random samples (n=500 each) of the Finnish population aged >16 years were drawn from the National Population Register. In the stratified sampling the elderly (aged ≥65) were over-represented to compensate for their lower absolute number in population and a possibly higher non-response rate. The content of the questionnaire for each sample was:

Sample 1 Background data (age, gender, education, whether experienced serious illness self, in family and when treating others, whether a person has at present an illness or impairment and its duration), tasks 1, 2 and 9, 15D questionnaire.

Sample 2 Background data, tasks 1, 4 and 9, 15D questionnaire.

Sample 3 Background data, tasks 6, 3 and 9, 15D questionnaire.

Sample 4 Background data, task 7, 5 and 9, 15D questionnaire.

Sample 5 Background data, the EuroQol descriptive system, 15D questionnaire, tasks 8 and 9.

At the end, the subjects assessed the time required to complete the questionnaire and how difficult it was on a four-point ordinal scale.

The valuation results are described by means and their 95% confidence intervals. Differences in means between samples are tested by Tukey multiple comparison tests in oneway ANOVA. For all analyses, the final samples were made comparable and compatible with the age and gender structure of the whole adult population (Statistics Finland 1993) by appropriate weighting.

Evaluation criteria and theoretical evaluations

The valuation methods and the resulting alternative value components are evaluated theoretically and empirically against four main criteria: feasibility, logical consistency, reliability and validity. The theoretical evaluations are presented here, the empirical ones in section 4.

Feasibility is judged empirically by measurement burden in terms of fill-in time, perceived difficulty as well as response and completion rates.
Consistency refers to the extent to which the valuations are consistent with a logical ordering of the health states in terms of goodness or badness (Dolan et al. 1993, Gudex et al. 1993). If a health state is logically better than another, its value should be higher. There is no logical order of importance between dimensions, but 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 subjects, who assigned a set of values consistent with that order. A logical order between these states and states `unconscious' and `dead' is less obvious: in this respect some people may

`legitimately' have `strange' values. It is expected though that most people regard the order on the questionnaire as logical and assign values accordingly.

The reliability of the measurement scores is concerned with the degree to which they can be repeated (McDowell and Newell 1987). It is estimated by taking repeated measurements and determining the agreement between them. The purpose with the 15D is not to derive new values each time the measure is used with the same or different individuals, but to derive a standard set of values once and then to apply them in all subsequent applications. Therefore it is not the test-retest repeatability over time at the individual level, but the repeatability and stability of valuations at the group or `social' level that is of a prime importance.

The repeatability of importance weights from the top of the scales was examined by comparing the results of an identical task 1 in samples 1 and 2, and from the bottom of the scales by comparing the results of task 6 in sample 3 and task 7 in sample 4. Since the duration specification was different in tasks 6 and 7, it is expected that the agreement may not as good as that obtained with identical tasks for the top of the scale importance weight. Oneway analysis of variance (with Tukey multiple comparison tests), Pearson correlation coefficient and Spearman rank correlation coefficient between the averaged sets of importance weights were used for analysis. Further, one averaged set was regressed on the other. If the regression coefficient deviates from 1, the constant term from 0, or the fit is poor, the sets do not agree.

Similar methods were used to examine the repeatability of level values. However, values for the levels were not derived in exactly the same way in two samples, since the duration specifications differed. This may complicate conclusions concerning repeatability, but shows on the other hand, whether duration is an issue (see below).

Validity indicates the extent to which accurate inferences can be made based on a measure. Validation is a process of hypothesis testing, by which the degree of confidence that can be placed in the inferences to be drawn from scores is determined (Streiner and Norman 1989). Since there is no gold standard for valuing health states, several types of validity have to be explored.

Content validity refers to how adequately the content of the measure reflects its aims. The question is: do all the aspects appear relevant to the concept being measured and are all aspects covered. The better the coverage, the more accurate and broader inferences can be validly drawn about the person or group under a variety of conditions and circumstances (McDowell & Newell 1987, Streiner & Norman 1989).

There are no absolute standards for judging content validity. However, a good starting point is to recall the aims of 15D. The 15D was developed for use in several areas, primarily though for measuring the effectiveness of health care programs in their economic evaluation, that is, in cost-utility analysis (Sintonen 1994). Richardson (1994) suggests three criteria relating to content validity, that a unit of measurement should meet, if used in that context.

First, there should be a broad concensus that the **units correspond to a socially desired outcome**. There is a wide consensus that health programs should be evaluated in terms of their contribution to length and quality of life. The measure of effectiveness should therefore capture, ideally in a single numerical score, the total change in health, gain or loss, incorporating both the change in the length and quality of life. A measure combining both aspects in a single score is quality-adjusted life years (QALYs).

However, in addition to gains in length and quality of life societies may and do seek to achieve also other objectives through their health care programs and systems, eg distributional objectives regarding access, health and income. Therefore a question arises whether these should be incorporated in the unit of measurement as well to make it compatible with the `socially desired outcome' or establish the `social value' of the outcome (Nord et al. 1993).

Here I agree with Broome (1993) that fairness and other distributional considerations must be considered separately although I have suggested an approach to combining the effects of health care programs in terms of QALYs, income and distribution (Sintonen 1981a). There are several reasons for that. All important value judgments affecting decisions are not known. Even if they were, trying to incorporate them is technically extremely difficult. It would only result in a unit of measurement that is by far more obscure and intractable than the already complex QALY concept is. This would violate Richardson's criterion 2 (see below) and alienate decision makers from economic evaluations. Economic evaluations can never be comprehensive enough to capture all factors at play in resource allocation decision making so as to `dictate' or `determine' decisions. At best they can assist decisions by producing information about the program costs and benefits as QALYs. Of course decision makers have to be aware of any `hidden' value judgments built-in in QALYs. Apart from that it should be left for them to decide what is the value of that information compared with other factors affecting decisions. The conclusion thus is that it suffices for the value component, if it reflects reasonably well not just the relative quality of various health states, but also a **trade-off between length and quality of life**.

This brings us to Richardson's criteria 2 and 3: the unit should have a **clear and unambiguous meaning** and a **meaningful interval property** to permit the summation of benefits. Since these criteria are closely related, they are looked at simultaneously. The question here is to what extent the QALYs obtained by using alternative valuation methods meet these criteria.

Most basic valuations were derived here by using a mixture of an RS (a 0-100 scale with clearly defined end points) and a ME (instructions and arrow prompts attracting ratio scale responses). Torrance (1986) claims that all the methods mentioned in section 1 produce at least an interval scale, ME a ratio scale. Nord (1991a) observed that even when unprompted, most respondents thought cardinally in terms of `percentages of fitness' when using an RS. This suggests that people tend to think `naturally' of the goodness of health states in the way values were elicited in this study. Broome (1993) argues that an RS of the type used here produces an appropriate cardinal `quality adjustment factor' for QALYs, whereas SG and TTO do not for two reasons. First, it is a `conventional wisdom' that preference methods (like the SG, TTO and PTO, which are based on locating indifference points) cannot give a cardinal scale for `good'. Second, the SG is valid only if the person is risk neutral, and the TTO only if the person does not discount. If these condition do not hold, and there is evidence that they do not (Mehrez and Gafni 1987, Olsen 1993), the SG or TTO do not produce a correct quality adjustment factor at all, let alone one that would indicate `good' cardinally.

In contrast to these arguments Richardson (1994) concludes that neither the RS nor the original, open-ended ME meet criteria 2 and 3 properly. Presumably then a mixture of RS and ME would not meet them, either. The same applies to the SG, whereas the TTO and PTO satisfy them.

In the face of these completely conflicting arguments it is impossible to say anything definitive about how well our valuations meet criteria 2 and 3. Moreover, there is no point in going deeper

into this issue here, since the SG, TTO and PTO were ruled out as unfeasible anyway. In the absence of agreement it will be assumed that our value components possess a reasonable cardinal property.

There are further criteria that may shed light on the pros and cons of the valuation approach used here compared with other methods (although not feasible in this context).

Broome (1993) argues that the quality weights should reflect person's **subjective good**. It is a basic tenet of welfare economics that the person is the best judge of his well-being. Also Rawls (1971) maintains that the person alone knows his plan of life and thus what is good for him. This would suggest that a valuation method, which elicits subjective preferences by asking people to think of themselves in the health states being valued (RS, ME, SG, TTO), is valid.

Some studies have shown that information about the **duration** of a state affects its valuation (Sackett and Torrance 1978). This has led to a requirement that when valuing health states their duration, and indeed a **certain prognosis** should be specified (Torrance 1986). This means that valuation should take place under conditions of complete certainty. On the other hand it has been suggested that valuation should be carried out **under the risk and uncertainty** associated with the interventions being evaluated (Loomes and McKenzie 1989, Richardson 1994).

Prognosis in essence contains two elements: the transition probabilities from one state to another (risk) and the duration of each state. More often than not, the prognosis is uncertain in a `genuine' sense, ie the probabilities and durations are unknown. Yet when using the SG, TTO or PTO, the occurrence and duration of the states to be valued are specified with certainty and usually with the RS and ME, a standard duration is given as well. These features do not conform with realism and therefore detract from the content validity of these methods.

In this study the valuations are based on a more realistic point of departure: the occurrence probability and duration of the states are left unspecified and thus uncertain as they mostly are in practice. The respondents can interpret them according to their attitude to uncertainty behind a veil of ignorance. The mean values obtained may thus reflect the average attitude to uncertainty in the population, which is an important feature of content validity. This view is also shared by Nord (1994). It can be questioned though whether such a complex phenomenon as risk and uncertainty can be fully taken into account in this simple way. However, to check empirically whether a specification of duration makes a difference, the duration was defined as one month and one year in some samples. The existence of possible differences in mean valuation between samples with different duration specifications was tested by using the Tukey multiple comparison tests with one-way ANOVA.

Construct validation involves gathering external empirical evidence, convergent or discriminant, so that meaningful inferences can be made with the measure. 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. Discriminant validity implies that the measure should not correlate with dissimilar, unrelated variables or measures (Streiner and Norman 1989).

To exhibit convergent evidence, the values produced by models (1)-(6) for the respondents' own health states were correlated (Pearson) with two other sets of values for those states. First, the respondents' own overall assessment of their health status (OVER) from task 9 (transformed to a 0-1 scale). Second, EuroQol values (EQ) for the respondents of sample 5 (n=359). These are based on the predicted values for the EuroQol states generated by Dolan and Kind (1993) with

regression analysis (main effect model) from the direct valuations of 42 states on a 0-100 visual analogue scale (EuroQol Group 1990) in an extensive Finnish postal survey with samples similar to those used in this study. In addition, regression techniques were used to explore the relationship between the value sets more closely. These analyses are based on sample 5.

Finally, another rough test of construct validity was carried out by applying a mapping procedure suggested by Nord et al. (1993). They chose four (basically EuroQol) health states (described in their table 2), which had been valued by the EuroQol valuation method in several countries and by the PTO in Norway and Australia. Then they mapped the states into three other health state descriptive and value systems and compared the values obtained by the states thereby. The systems considered were Quality of Well-Being scale (QWB) (Kaplan and Anderson 1988), the McMaster Health Classification System (Torrance et al. 1982) and the Rosser/Kind index (R/K) (Rosser and Kind 1978). Here this comparison was extended so that the same group (E.N. and J.R.) and the author mapped the same states independently into the 15D.2 descriptive system and these mappings were then assigned a value by models (1)-(6).

4 Results

Feasibility: Table 1 shows statistics relating to feasibility. For comparison, the samples were rendered compatible with the population age and gender structure. In samples 1-3 with comparable questionnaires, the response rate was 43-46%. In sample 4, where states 'unconscious' and 'dead' were not included for valuation the response rate was 52%. Sample 5 had only one valuation task, resulting in a clearly higher response rate. The completion rates for importance weight were 71-82% and slightly lower for level values, especially for state 'dead', reflecting the well-known difficulty with valuing that state. The average fill-in time was about half an hour (in sample 5 only 15 minutes). The average difficulty with filling in the questionnaire was about 3, that is, "relatively easy".

Table 1

Descriptive statistics relating to the samples and the feasibility of their tasks

Sample	Response rate %	Mean age	Male %	Completion rate for I_j	Completion rates for levels		Mean fill-in time, min	Mean difficulty‡
					Level 2	`Being dead'		
1	43	46	48	71-76	64-74	61-66	36	2.6
2	46	45	50	78-82	69-79	55-60	34	2.7
3	45	47	48	75-77	67-72	59-61	27	2.7

4	52	45	48	78-82	70-77	†70-74	29	2.7
5	72	45	48	n.a.	n.a.	61	15	3.1

Notes † Level 5.
‡ On a 1-4 scale, where 1 = very difficult and 4 = very easy.

Logical consistency. The subjects valued the five levels within each dimension quite consistently. Depending on the dimension and level, there were only 0-2.5% of respondents in samples 1-3, who assigned a lower value to a logically better level. As expected, the percentages were higher with states `unconscious' and `dead': 1-11.6% assigned a higher value to unconsciousness than to level 5, and 18-20% a higher value to being dead than to unconsciousness. However, no observations were dropped due to `inconsistency'.

Reliability. The mean I_j weights and their 95% confidence intervals from samples 1-4 are depicted in table 2. The Tukey multiple comparison tests with oneway ANOVA showed that pairwise none of the mean I_j weights (from the top levels) differed significantly between samples 1 and 2. The Pearson correlation coefficient between the two averaged sets of weights was .970 and the Spearman rank correlation .963. When the set from sample 1 (I_{j1}) was regressed on the set from sample 2 (I_{j2}), the equation obtained was $I_{j1} = -.001 + 1.014I_{j2}$, R^2=.940. The constant term did not deviate significantly from zero (t=.20, df 13) and the regression coefficient from 1 (t=.20, df 13). These statistics indicate that the two sets agree quite well and the reliability at the group level is thus good. For `final' I_j weights from the top of the dimensions (column 5 in table 2) the samples were pooled.

Pairwise there were no significantly different mean I_{jb} weights (from the bottom levels) between samples 3 and 4, either (table 2). The Pearson correlation between the two averaged sets was .977 and the Spearman rank correlation .938. When the set from sample 3 (I_{jb3}) was regressed on the set from sample 4 (I_{jb4}), the equation was $I_{jb3} = .005 + .918I_{jb4}$, R^2=.954. The constant term was not significantly different from zero (t=1.39, df 13) and the regression coefficient from 1 (t=1.46, df 13). The two sets agree thus quite well and the reliability at the group level is good. This agreement also suggests that it does not make a difference to the weights whether the duration is unspecified or defined as 1 year. Therefore for `final' I_{jb} weights from the bottom of the dimensions (column 6 in table 2) samples 3 and 4 were pooled.

In pairwise comparisons, most of the mean I_j weights from the top levels differed significantly from those from the bottom levels (table 2).

For each dimension, the Pearson correlation between the three sets of averaged level values from samples 1-3 was $\approx$ .99. However, this does not necessarily imply a good agreement between the sets (Bland and Altman 1986). When level values were compared pairwise between the samples, samples 1 and 2 did not differ significantly at any level of any dimension in the 90 comparisons carried out so the averaged sets from these samples agree quite well. In 41 comparisons there was no significant difference between the three samples. In 24 comparisons, sample 2 deviated from sample 1, but not from sample 3. In the remaining 25 comparisons, sample 3 deviated both from sample 1 and 2. There is thus a tendency that the level values in sample 3 with the duration of states defined as one year are higher than in samples 1 (unspecified duration) and 2 (one month duration). For final level values, samples 1 and 2 were pooled.

Table 2

The mean l_j weights for the dimensions from different samples, and the `final' mean l_j weights from pooled samples (95% confidence intervals in parentheses)

Dimension	l_{j1} (top) Sample 1 n = 137	l_{j2} (top) Sample 2 n = 167	l_{jb3} (bottom) Sample 3 n = 158	l_{jb4} (bottom) Sample 4 n = 195	Final l_j Sample 1+2 n = 304	Final l_{jb} Sample 3+4 n = 353
Sleeping	.070 (.067-.073)	.069 (.066-.072)	.079 (.068-.089)	.085 (.076-0.94)	.070 (.067-.072)	.082 (.075-.089)
Breathing	.083 (.079-.087)	.084 (.082-.087)	.074 (.065-.083)	.073 (.065-.081)	.084 (.082-.086)	.074 (.068-.080)
Eating	.071 (.067-.074)	.071 (.068-.074)	.039 (.034-.044)	.043 (.038-.048)	.071 (.068-.073)	.041 (.039-.045)
Speech	.066 (.063-.070)	.066 (.063-.069)	.058 (.050-.065)	.067 (.059-.075)	.066 (.064-.069)	.063 (.057-.069)
Mental function	.086 (.083 - .089)	.085 (.081-.088)	.044 (.032-.055)	.042 (.036-.047)	.085 (.083-.088)	.042 (.036-.049)
Mobility	.074 (.070-.078)	.068 (.065-.070)	.036 (.030-.042)	.033 (.029-.038)	.070 (.068-.073)	.035 (.031-.038)
Discomfort/symptoms	.063 (.059-.067)	.062 (.058-.066)	.045 (.040-.051)	.041 (.037-.046)	.062 (.060-.065)	.043 (.040-.047)
Sexual activity	.052 (.048-.056)	.057 (.053-.061)	.097 (.082-.112)	.098 (.090-.106)	.054 (.052-.057)	.097 (.089-.105)
Hearing	.059 (.056-.063)	.059 (.056-.062)	.101 (.092-.109)	.109 (.101-.118)	.059 (.057-.061)	.106 (.099-.112)
Vitality	.073 (.069-.078)	.077 (.073-.082)	.084 (.079-.090)	.080 (.074-.086)	.076 (.072-.079)	.082 (.079-.086)
Distress	.061 (.057-.065)	.061 (.058-.065)	.082 (.076-.089)	.079 (.073-.084)	.061 (.059-.064)	.080 (.076-.085)
Usual activities	.076 (.073-.080)	.076 (.072-.080)	.064 (.059-.070)	.058 (.054-.062)	.076 (.073-.079)	.061 (.057-.064)
Elimination	.063 (.059-.066)	.061 (.058-.064)	.044 (.038-.049)	.042 (.038-.047)	.062 (.059-.064)	.043 (.039-.046)
Depression	.051 (.047-.056)	.052 (.049-.056)	.081 (.074-.088)	.079 (.073-.085)	.052 (.049-.055)	.080 (.075-.084)
Vision	.052 (.047-.056)	.052 (.048-.056)	.072 (.065-.079)	.071 (.064-.078)	.052 (.049-.055)	.071 (.067-.076)
S	1.000	1.000	1.000	1.000	1.000	1.000

It turned out that $\Sigma c_j>1$ and $\Sigma d_j>1$. The value of c and d was -.9999. The value of the worst combination of levels from task 8 was -.334 on a 0-1 scale (healthy=1, dead=0). This suggests a scaling factor of 1.5858 for model (3) and 1.334 for model (4). These imply that the utility of the worst combination in model (3) is -.5858 and the value of that combination is -.334 in model (4). In models (5) and (6) this state was assigned a utility/value of .1, which is the same as the value of that state in models (1) and (2).

Validity. Table 3 shows descriptive statistics and correlations of the values for the respondents of sample 5 on a 0-1 scale (0=dead) as produced by the different valuation models. The mean represents the average HRQOL score of the adult Finnish population. Models (1) and (2) generate a score of .92. The EuroQol scores .84, model (5) .78 and the average overall assessment is .76. The remaining models yield much lower scores. Indeed, models (3) and (4) suggest that 12-25% of the adult Finnish population would live in a state worse than death. According to model (6) about 58% of adult population would be in a state with an HRQOL score less than .51, ie the minimum for models (1) and (2).

The scores generated by models (1) and (2) correlate strongly, almost completely, with each other and also highly with those from the other models. The scores from model (1) correlate slightly higher than those from model (2) with the other scores apart from EuroQol and overall assessment. The EuroQol scores correlate highest with the model (2) scores. The overall assessments have the highest correlation with the EuroQol and second highest with model (2). These findings provide solid convergent evidence of construct validity for the `original' 15D value components based on models (1) and (2). Below are the best regression equations for converting the scores from the other models into model (2).

$v_{HM2} = .016+.985v_{HM1}$, $R^2=.992$

$v_{HM2} = .820+.258\ln(u_{HM3}+1)$, $R^2=.943$

$v_{HM2} = .865+.225\ln(v_{HM4}+1)$, $R^2=.787$

$v_{HM2} = .659+.339u_{HM5}$, $R^2=.855$

$v_{HM2} = 1.018+.095\ln(v_{HM6})$, $R^2=.742$

$v_{HM2} = .456+.805EQ-.291(EQ)^2$, $R^2=.612$

$v_{HM2} = .707+.281OVER$, $R^2=.377$

Table 3

Descriptive statistics and correlations of the values (quality of life scores)
for the respondents of sample 5 on a 0-1 scale (0=dead)

Variable	Mean	SD	Min	Max	N	% < .51	% < 0
V_{HM1}	.92	.09	.51	1.00	329	0.00	0.00
V_{HM2}	.92	.09	.51	1.00	329	0.00	0.00
U_{HM3}	.56	.41	-.57	1.00	329	37.1	12.5
V_{HM4}	.37	.44	-.33	1.00	329	58.7	24.9
U_{HM5}	.78	.24	.10	1.00	329	14.3	0.00
V_{HM6}	.49	.33	.10	1.00	329	57.8	0.00
EQ	.84	.20	.02	1.00	342	9.9	0.00
OVER	.76	.19	.15	1.00	336	15.2	0.00

Correlations

	V_{HM1}	V_{HM2}	U_{HM3}	V_{HM4}	V_{HM5}	V_{HM6}	EQ	OVER
V_{HM1}	1.000							
V_{HM2}	.996	1.000						
U_{HM3}	.950	.941	1.000					
V_{HM4}	.841	.830	.939	1.000				
V_{HM5}	.925	.920	.982	.874	1.000			
V_{HM6}	.762	.751	.879	.987	.806	1.000		
EQ	.755	.761	.746	.681	.727	.626	1.000	
OVER	.600	.605	.594	.530	.586	.486	.643	1.000

Notes All correlations significant at a .001 level

The results of another test of construct validity are in table 4, which shows the values for four health states mapped into four different health state descriptive systems and ten valuations

methods. The very low, mainly negative values generated by models (3) and (4) appear also here. Model (6) yields values that are all very close to the value of the worst combination and can hardly discriminate between the states. The situation is slightly better in this respect with model (5), but it still produces a set of very low values as does the McMaster (underlying model (3)) and EuroQol. The highest values are generated by the PTO, Rosser/Kind and model (2): these are roughly of the same magnitude and relatively close to each other.

Table 4

Comparison of values (range midpoints) for four health states from ten valuation methods on a 0-1 scale (0=dead)

State	EQ[1]	PTO[1]	QWB[1]	McMaster[1]	R/K[1]	M2	M3	M4	M5	M6
A	.23	.70	.52	.24	.50	.75	-.40	-.29	.13	.10
B	.33	.65	.59	.36	.70	.80	-.32	-.26	.15	.11
W	.60	.90	.68	.35	.94	.86	.08	-.12	.36	.13
Z	.20	.78	.50	.08	.68	.77	-.41	-.29	.12	.10

[1] Based on tables 5 and 6 by Nord et al. (1993)

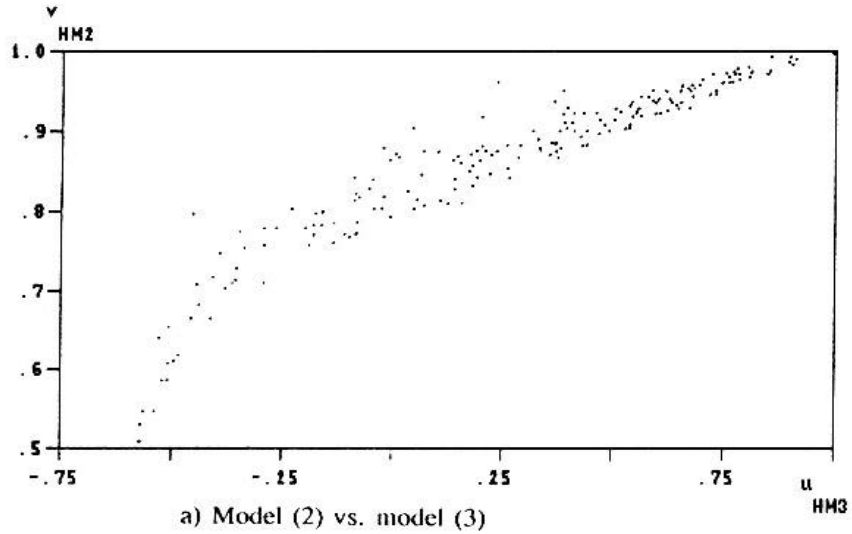
The findings in tables 3 and 4 suggest that models (3)-(6) compress the values strongly towards the lower end of the range. This can also be seen clearly from figure 1, where the model (2) scores for the respondents are plotted against those from models (3)-(6). The compression is particularly striking with models (4) and (6), but is also marked in model (5). This phenomenon implies that with poorer health states, the models loose their sensitivity, that is, they are not able to discriminate valuewise between the states. This can be illustrated with an example.

Take the worst case in sample 5, who scored .505 in model (2). His score was -.570, -.333, .099 and .099 in models (3)-(6), respectively. Improve the state by one level on each dimension where possible (not on three). This should represent a substantial improvement. In model (2) the improvement was .183, and in models (3)-(6) .120, .014, .025 and 0, respectively. Thus models (4) and (5) indicated a marginal improvement, but model (6) nothing at all. Now deteriorate the state by one level on each dimension where possible (not on three). This should be a notable worsening in the state. Indeed, in model (2) the score declined by .179, but only by .013 in model (3) and by 0 in models (4)-(6). In models (3)-(6) there is a `floor' effect of having no value to go `down'. They suffer from a serious lack of sensitivity in terms of discriminatory power and responsiveness to change already with relatively good health states (as indicated by overall assessment, EuroQol and models (1) and (2)).

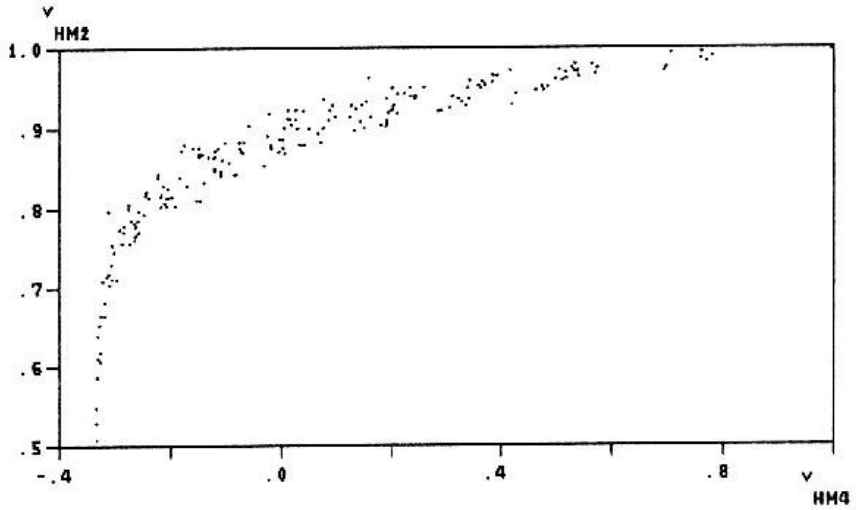
5 Discussion

This study has shown that it is possible to derive reliable and valid valuations for the 15D health state descriptive system by using self-administered questionnaires in a postal survey.

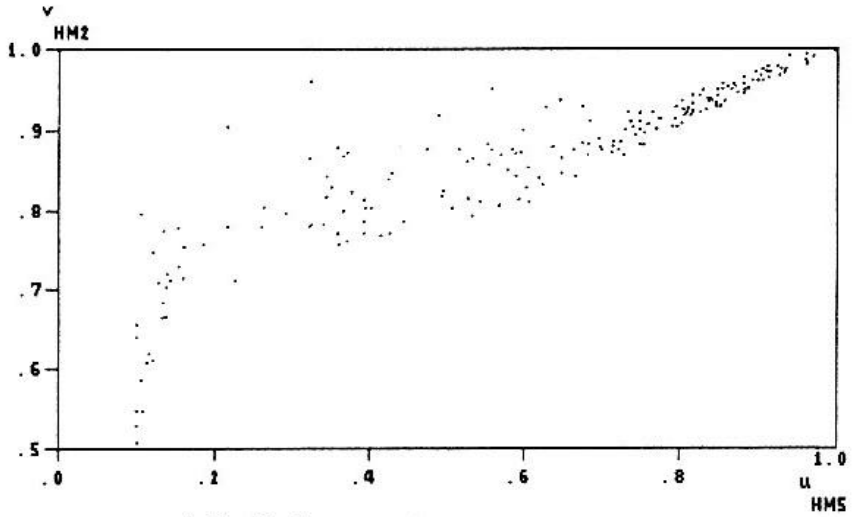
Fig. 1. The model (2) scores for the respondents plotted against those from models (3)-(6)



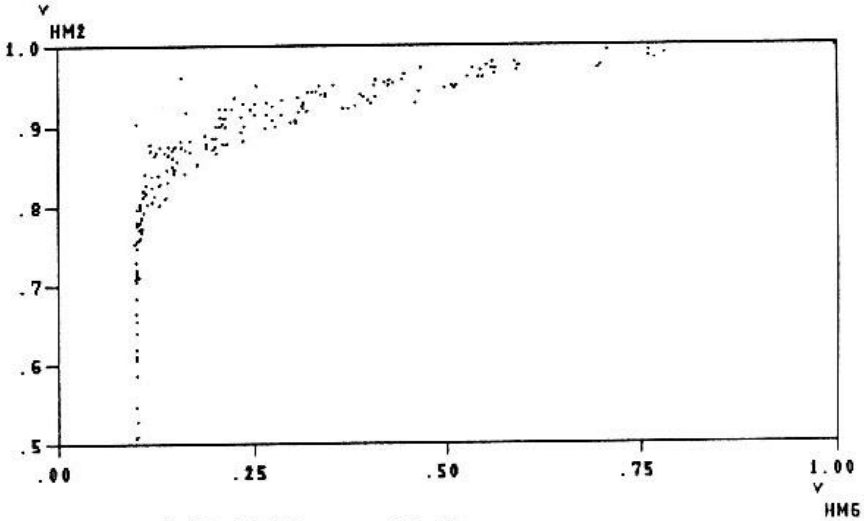
a) Model (2) vs. model (3)



b) Model (2) vs. model (4)



c) Model (2) vs. model (5)



d) Model (2) vs. model (6)

In samples 1-3, which are the most important ones in this study, the response rate was 43-46% and completion rate for valuations about 70%. This means that effectively about one third of the original sample was usable in the analyses. However, with appropriate weighting the final samples could be rendered fully compatible with the age and gender structure of the Finnish adult population. It is likely that much higher response and completion rates would be achievable by interviewing or at least by having someone to introduce and motivate the respondents to the valuation tasks.

Since there is no gold standard the validity of health state valuations for the 15D health state descriptive system can never be proven. In this situation Nord (1992) suggests that the criterion test of validity may consist of testing for reflective equilibrium, ie examining to what extent preference statements that are inferred from health state valuations are in accordance with preferences that are directly elicited. Essentially this comes down to saying that the real proof of the pudding lies in the eating: the question is about how plausible the values are in the context where they are used.

It is highly implausible that 12-25% of adult population would be in health states worse than death as the scores generated by models (3) and (4) suggest, ie that percentage would rather be dead than go on living in their present health states. This makes these scores invalid for inferences about people's HRQOL cross-sectionally. They might still be useful in indicating changes in HRQOL over time, but a considerable compression of values toward the lower limit makes their validity questionable also in this respect. It is also highly unlikely that health states two levels apart on 9 dimensions and one level apart on the rest of 15 dimensions would be virtually or exactly equal in HRQOL as the scores from models (4)-(6) suggest. This property renders these scores invalid both for discriminating between the HRQOL in different health states and for inferences about changes in HRQOL over time.

Why do not models (3)-(6) work, at least with the 15D health state descriptive system in spite of their `theoretical soundness' in the light of the MAU theory? There are undeniably methodological differences in how values were derived in this study and by Torrance et al. (1982, 1992). When similar rating scales were used, Torrance et al. defined the states as lasting for lifetime, whereas in this study the duration was unspecified or much shorter (max one year). However, duration did not appear to make much difference in valuations. This is consistent with results from EuroQol work, where durations up to 10 years did not affect the valuations (Ohinmaa and Sintonen 1994). Second, when Torrance et al. (1982) valued corner states, they asked respondents to imagine being at the lowest level on one dimension and at the best on the rest. In the 1992 study the procedure was basically the same, but there were some modifications to make the states to be valued less unrealistic. In this study the respondents were simply requested to think of one dimension at a time.

These differences may have some effect on results, but a major reason for the features models (3)-(6) exhibit probably is that with 15 dimensions the model is not appropriate. In the studies of Torrance et al. and in this study, the value of the interaction parameter c (or d) has turned out very close to -1. Therefore the $c_j u_j(x_j)$ values have to be on average at least .55, before their product with 15 dimensions exceeds 0 (with four decimal point precision). For example in model (5) this value was typically between levels 3 and 4 and in model (6) between levels 2 and 3. This roughly implies that a health state composed of middle levels 3 would already be regarded as equivalent to being dead, that is, people would give up all their remaining life. This is not very plausible, but it explains the very strong compression of values toward the low end of the valuation space. With fewer dimensions this property is less marked, but as table 4 shows even with four dimensions the

multiplicative model yields quite low values (McMaster). The conversion factors to utility alleviate the compression, but it is not yet clear, what the `right' factor is. A multilinear model would not be so constrained on interactions, but with more than four dimensions, its estimation becomes unfeasible (von Winterfeldt and Edwards 1986).

This leaves us with the additive models (1) and (2), which in the light of the diagnostic statistics ($\Sigma c_j>1$ and $\Sigma d_j>1$) are not `correct' ones. One could argue that if the above models do not work, a holistic valuation approach is needed to account for interactions between dimensions and thus to derive valid valuations. The problem is that with 15 dimensions, holistic valuation is not feasible. However, the validity of scores from models (1)-(2) can be tested against two sets of holistic valuations: the respondents' EuroQol score and own overall rating of health status, which is the most holistic valuation of all.

Apart from model (5) the average scores from models (1) and (2) come closest in magnitude to these holistic scores. They also have highest correlations with the holistic valuations among the models considered. Yet the difference between the average model (1)-(2) scores and overall assessments is not negligible and the correlations (.600-.605) are the lowest that model (1)-(2) scores have with other sets of values. There may be several reasons for that. The overall assessment may capture health problems that are not covered by the 15D descriptive system, in which case it would be a more valid measure. On the other hand it may be affected by clinical medical information, which may have no effect on the 15D valuations and is not captured (on purpose) by the generic 15D descriptive system, or by extraneous and situational, non-health-related determinants. It may also be that the RS in the overall rating encourages people to take middle values. This unstandardised, black-box nature of overall rating makes it an inadequate criterion for the ideal RHQOL score. Anyway, the additive models show substantial validity and robustness against holistic valuations.

There are two further reflective equilibrium tests for the plausibility and thus validity of the scores generated by models (1) and (2) against those by the other models in QALY calculations. One is a time trade-off type of test. Take the average adult citizen of 48 years of age in the sample. Her/his HRQOL is .76, .84 or .92 according to overall rating, EuroQol or models (1)-(2), respectively. Assume that her/his expected remaining length of life at that level of HRQOL would be 30 years, which would yield her/him 22.8, 25.2 or 27.6 QALYs, i.e, years in full HRQOL. This implies that if the citizen were promised that many years in full HRQOL, (s)he would be willing to sacrifice 7.2, 4.8 or 2.4 years at her/his present level of HRQOL (overlooking the problem of discounting) and would be equally well-off. To my knowledge trade-offs like these have not been elicited in practice, but my guess is that our average citizen's willingness to sacrifice would be closer to the last figure than the others. Should that be the case it would suggest that the values generated by models (1) and (2) provide a more realistic trade-off between length and quality of life, and thus a property that is necessary for valid QALY results. I leave it to the reader to carry out her/his own tests and decide, which set of values is most valid in this respect.

The other is a PTO test in table 3, advocated mainly by Nord (1992, 1994). With reference to table 3 Nord et al. (1993) concluded that the EQ, QWB and McMaster values are far too low to be plausible and valid for social choice involving a trade-off between length and quality of life, whereas the Rosser/Kind values appear quite reasonable, since they come closest to the PTO values serving as the criterion for this conclusion. Having a look at the model (2) values and extending the logic it can be inferred that they come on average equally close to the PTO values as the Rosser/Kind values. In this sense they appear to satisfy the test of plausibility quite well and clearly better than the other methods considered in this paper.

For several reasons these results should not be taken as indicative that the PTO values show the `right' trade-off between length and quality of live, let alone the whole `social value' of outcome. The mappings are inevitably inexact. The PTO method involves a lot of unresolved theoretical and practical problems (Nord 1994). In the study underlying table 3 (Nord et al. 1993) states A-Z were assessed against emergency life saving (saving from dying and restoring full health). Emergency life saving may assume virtually a lexicographical priority over other health programs and therefore it may not be the most appropriate point of reference. As discussed earlier, it is impossible to establish the `social value' of an outcome in advance. However, the PTO method provides a useful framework of making the length-quality trade-off explicit and allowing thus reflections about how plausible the trade-offs implicit in different sets of values are. Again the reader can carry out tests of her/his own, and as I expect, come consequently to a conclusion that the model (1) and (2) scores appear quite plausible.

So what is the explanation for models (1) and (2) performing so well in these tests? It may be that the additive model provides after all a reasonable approximation to how people value multi-attribute health states. Further research is needed to shed light on this question and to increase our understanding of valuation processes. It may be that through interaction of different design factors the models produce values that consistently perform well on a test reflective equilibrium. If that is the case and if such goodness of fit is empirically demonstrated as it here, the model should, as Nord (1992) points out, be regarded as a valid approach to eliciting health state values.

Dolan and Kind (1993) report that the predicted values for EuroQol states are often counter-intuitive, ie logically worse states get higher scores. However, it must be borne in mind that the EuroQol values used here may not be the `final' ones. A search for the `best' method of modelling and predicting EuroQol values is still under way so eventually a method may be developed that produces consistent values. Anyway, validity-detracting inconsistencies are not possible with the scores provided by models (1) and (2). They are both consistent and sensitive over the whole valuation space in the sense that if a health state in the 15D descriptive system is logically better than another, its score will be higher, and vice versa. Thus these value components meet Richardson's (1994) fourth criterion for the unit of measurement in the QALY context: the unit should be sensitive to variation in the relative dimensions of the outcome.

Models (1) and (2) generate only positive scores, ie all health states are thus regarded as preferable to death. Of course it is possible to scale the scores so that also negative ones are allowed. However, staying on the positive side is consistent with the ethical climate in most societies: even if the individual would consider her/his state worse than death and even if most people would regard it as such, legislation does not acknowledge such states and allow people in those states be helped to die for improving their quality of life. Thus societies assign a positive score to all health states (except brain death perhaps) anyway in their health policy. From this point of view nothing is gained by allowing negative scores, but considerable analytical and ethical complexities are created by doing so. Contrary to their 1982 paper also Torrance et al. (1992) restrict the scores to the positive area.

It is very likely that the values generated by models (1) and (2) with Finnish data are widely generalisable and valid in Western-type countries. One of the most important lessons from the EuroQol project has been that the health state valuations elicited in the same way from the general public in England, Finland, Norway, The Netherlands and Sweden are strikingly similar (EuroQol Group 1990, Björk 1992, Sintonen 1993). The same has been observed also with the NHP valuations in England, Sweden and Finland (Hunt and Wiklund 1987, Koivukangas et al. 1992).

This means that to apply the 15D, it is not necessary to engage first in a laborious and resource consuming valuation of the instrument (although it is welcomed and encouraged for comparative purposes), but the users can quite confidently use the valuations generated here. All that is required is that the respondents fill in the 15D questionnaire (health state descriptive system), which takes a few minutes, the responses are entered in a computer and combined with the value component to produce the 15D scores.

Computationally it is equally simple to use either model (1) or (2) scores. Personally I prefer model (2) scores for theoretical reasons and due to the fact these scores have a slightly higher correlation with the EuroQol scores and overall rating. Should someover want to carry out the valuations in a patient or population group with minimal effort, model (1) is simpler since it involves one stage less than model (2), but still yields values very close to those from model (2).

In sum, the 15D is probably the most sensitive and comprehensive HRQOL measure presently available that combines the advantages of a profile and single index score measure. In both capacities it exhibits strong evidence for reliability and validity. At this level of sensitivity, it seems to provide a more plausible and valid set of health state valuations for QALY measurements in cost-utility analysis than any other measure.

The 15D is being used in numerous evaluation projects in Finland and in several other countries (Sintonen and Pekurinen 1993). It is hoped that this and an earlier paper (Sintonen 1994) would encourage reseachers with various research interests involving quality of life measurement to seriously consider including the 15D in their tool kit and give it a try.

REFERENCES

Björk S (ed.) EuroQol Conference Proceedings. Discussion Paper No 1. IHE Working Paper 1992:2, Lund 1992.

Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, Feb. 9, 307-310.

Broome J (1993) Qalys. Journal of Public Economics 50, 149-167.

Bullinger M (1993) Indices versus profiles - advantages and disadvantages. In Walker SR, Rosser RM (eds) Quality of life assessment. Key issues in the 1990s. Kluwer Academic Publishers, Dordrecht, 209-220.

Dolan P, Gudex C, Kind P, Williams A (1993) Valuing health states: A comparison of methods. Paper presented to Strathclyde meeting of the Health Economics Study Group, June 30-July 2, 1993.

Dolan P, Kind P (1993) Modelling the EuroQol data: Some preliminary results. Paper presented at the EuroQol Group Meeting, Rotterdam Oct. 1993.

EuroQol Group (1990) EuroQol: A new facility for the measurement of health-related quality of life. Health Policy 16, 199-208.

Froberg D, Kane R (1989) Methodology for measuring health-state preferences - II: Scaling Methods. International Journal of Epidemiology 42, 459-471.

Gudex C, Kind P, van Dalen H, Durand M-A, Morris J, Williams A (1993) Comparing scaling methods for health state valuations - Rosser revisited. University of York, Centre for Health Economics, Discussion Paper 107.

Hunt SM, Wiklund I (1987) Cross-cultural variation in the weighting of health statements: a comparison of English and Swedish valuations. Health Policy 8, 227-235.

Kaplan RM, Anderson JP (1988) A general health model: Update and application. Health Services Research 23, 203-235.

Keeney RL, Raiffa H (1976) Decisions with multiple objectives: Preferences and value trade-offs. Wiley, New York.

Koivukangas P, Koivukangas J, Ohinmaa A, Kivelä S-L, Krause K (1992) NHP - a method for measuring health-related quality of life in health services evaluation. Journal of Social Medicine 29, 229-235.

Loomes G, McKenzie L (1989) The use of QALYs in health care decision making. Social Science & Medicine 28, 299-308.

McDovell I, Newell C (1987) Measuring health: A guide to rating scales and questionnaires. Oxford University Press, New York, Oxford.

Mehrez A, Gafni A (1987) An empirical evaluation of two assessment methods of utility measurement for life years. Socio-Econonic Planning Science 21, 371-375.

Mehrez A, Gafni A (1989) Quality-adjusted life years, utility theory, and healthy-years equivalents. Medical Decision Making 9, 142-149.

Nord E (1992) Methods for quality adjustment of life years. Social Science & Medicine 34, 559-569.

Nord E (1994) The person trade-off approach to valuing health care programs. National Centre for Health Program Evaluation, Working Paper 38. Melbourne.

Nord E (1991) The validity of a visual analogue scale in determining social utility weights for health states. International Journal of Health Planning and Management 6, 234-242.

Nord E, Richardson J, Macarounas-Kirchmann K (1993) Social evaluation of health care versus personal evaluation of health states. Evidence on the validity of four health-state scaling methods using Norwegian and Australian surveys. International Journal of Technology Assessment in Health Care 9, 463-478.

Ohinmaa A, Sintonen H. (1994) The effect of duration on the values given to the EuroQol states. In Busschbach JJV, Bonsel GJ, de Charro FT (Eds) EuroQol Plenary Meeting, Rotterdam 1993, 6-8 October. Discussion Papers, Erasmus University, 51-59.

Olsen JA (1993) Time preferences for health gains: An empirical investigation. Health Ecomomics 2, 257-265.

Patrick DL, Bush JW, Chen MM (1973) Methods for measuring levels of well-being for a health status index. Health Services Research 8, 228-245.

Rawls J (1971) A theory of justice. Harvard University Press, Cambridge.

Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC (1984) Preferences for health outcomes. Comparison of assessment methods. Medical Decision Making 4, 315-329.

Richardson J (1994) Cost-utility analysis: What should be measured? Social Science & Medicine 39, 7-21.

Rosser R, Kind P (1978) A scale of valuations of states of illness: Is there a social consensus? International Journal of Epidemiology 7, 347-358.

Sackett DL, Torrance GW (1978) The utility of different health states as perceived by the general public. Journal of Chronic Diseases 31, 697-704.

Sintonen H (1994) 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.

Sintonen H (1981a) An approach to economic evaluation of actions for health. Official Statistics of Finland, Special Social Studies XXXII:74, Government Printing Centre.

Sintonen H (1981b) An approach to measuring and valuing health states. Social Science & Medicine 15C, 55-65.

Sintonen H (ed.) EuroQol Conference Proceedings. Helsinki, October 1992, Discussion Paper No 2, Kuopio University Publications E. Social Sciences 8. Kuopio University Printing Office 1993.

Sintonen H, Pekurinen M (1993) A fifteen-dimensional measure of health-related quality of life (15D) and its applications. In Walker SR, Rosser RM (eds) Quality of life assessment. Key issues in the 1990s. Kluwer Academic Publishers, Dordrecht, 185-195.

Statistics Finland (1993) Statistical yearbook of Finland 1993. Vol. 88. Printing Centre, Helsinki.

Streiner DL, Norman GR (1989) Health measurement scales: A practical guide to their development and use. Oxford University press, Oxford, New York, Tokyo.

Torrance GW (1986) Measurement of health state utilities for economic appraisal. Journal of Health Economics 5, 1-30.

Torrance GW (1976) Social preferences for health states: An empirical evaluation of three measurement techniques. Socio-Economic Planning Science 10, 129-136.

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

Torrance GW, Zhang Y, Feeny D, Furlong W, Barr R (1992) Multi-attribute preference functions for a comprehensive health status classification system. CHEPA Working Paper Series No. 92-18, McMaster University, Hamilton 1992.

Winterfeldt D von, Edwards W (1986) Decision analysis and behavioral research. Cambridge University Press, Cambridge 1986.