ELSEVIER

Journal of Health Economics 18 (1999) 551–571

JOURNAL OF HEALTH ECONOMICS

www.elsevier.nl/locate/econbase

Estimating parametric relationships between health description and health valuation with an application to the EuroQol EQ-5D

Jan J.V. Busschbach [a,*], Joseph McDonnell [a], Marie-Louise Essink-Bot [b], Ben A. van Hout [a]

[a] *Institute for Medical Technology Assessment (iMTA), Erasmus University Rotterdam, PO Box 1738, 3000 DR Rotterdam, The Netherlands*

[b] *Department of Public Health, Erasmus University Rotterdam, Rotterdam, The Netherlands*

Received 17 April 1996; received in revised form 8 September 1998; accepted 23 March 1999

Abstract

Generic health status measures classify patients into different health states. For example, the EQ-5D descriptive system developed by the EuroQol Group classifies patients into 243 health states. Empirical values for the health states are available for only a selection (mostly 12 to 45) of these health states. Several parametric relationships between the descriptive system and the known values can be formulated to estimate the values for the unrecorded health states. This paper describes several of these modeling exercises in a comprehensible way, using the EQ-5D as an illustration. It is shown that the estimation task does not depend on the meaning of the values, but does depend on the selection of the empirically valued health states and the assumptions about the relationship between these values and the descriptive system.

Keywords: EQ-5D; Health states; EuroQol

1. Introduction

Health state valuation measures like the EQ-5D, the Health Utility Index and the Quality of Well-Being are used to attribute a value to the health state of a

* Corresponding author. Tel.: +31-10-4088555; Fax: +31-10-4089092; E-mail: busschbach@bmg.eur.nl

0167-6296/99/$ - see front matter
PII: S0167-6296(99)00008-9

patient. These measurements typically first classify a patient into one of several possible health states. For example, the EQ-5D descriptive system, developed by the EuroQol Group can classify patients into one of 243 possible health states, plus the states *death* and *unconscious* (Brooks, 1996). Once the patient has been classified into such a health state, the researcher can assign a relevant value. These values are based on previous research where usually not all the values of the possible health states have been measured. Generally, empirical values are only available for a selection of all possible health states. By estimating a parametric relationship between the descriptive system and the known values however, it is possible to estimate the values for the unrecorded health states. In this article, we compare different approaches and categorise some basic methodological problems using the EQ-5D instrument.

One of the methodological starting points of the EuroQol Group was the assumption that health can be characterised by a set of scores applied to five aspects of health status. The EQ-5D refers to these aspects as 'dimensions'. The dimensions each comprise three levels: no problems, some/moderate problems and extreme problems/unable to (Table 1). Within the EQ-5D classification system, every individual health state can be described by a row vector $\vec{x}(x_1, x_2, \ldots, x_5)$ in which the element x_i represents the score on dimension i. Thus, $x_1 =$ the score on mobility, $x_2 =$ the score on self-care, etc. The score on a dimension is '1' if it is the highest level and '3' if it is the lowest (Table 1). For example, a health state such as: some problems in walking about, no problems with self-care, no problems with performing usual activities, moderate pain and

Table 1
The EuroQol health dimensions and their scores

Dimension	Levels	Scores
Mobility	No problems in walking about	1
	Some problems in walking about	2
	Confined to bed	3
Self-care	No problems with self-care	1
	Some problems with washing or dressing self	2
	Unable to wash or dress self	3
Usual activities	No problems with performing usual activities (e.g., work, study, housework, family or leisure activities)	1
	Some problems with performing usual activities	2
	Unable to perform usual activities	3
Pain/discomfort	No pain or discomfort	1
	Moderate pain or discomfort	2
	Extreme pain or discomfort	3
Anxiety/depression	Not anxious or depressed	1
	Moderately anxious or depressed	2
	Extremely anxious or depressed	3

moderate anxiety could be represented by the row vector (2,1,1,2,2), usually abbreviated to 21122. This notation is used throughout this paper. The states *death* and *unconscious* cannot be described within the classification system.

With five dimensions and three levels per dimension, there are $3^5 = 243$ possible health states. Ideally, respondents should value all 243 health states, but in practice they can value only approximately 30 health states. This means that it is an extremely laborious task to obtain empirical values for all the possible health states. For this reason, investigators use parametric models to estimate the values for the health states not included in the empirical valuation. The parametric model should predict the values for the health states on the basis of the scores of the descriptive system. In algebraic terms we seek a relationship as follows:

$$V = V(\vec{x}) \tag{1}$$

where V is the value of the health state, $\vec{x}$ is the vector of the health state description and $V(\vec{x})$ is the value function.

Because it is not possible to describe the states *death* and *unconscious* using this classification system, we cannot estimate their values using a parametric relationship. Therefore, the values from these states can only be determined empirically.

In this paper, we demonstrate how it is possible to estimate such value functions $V(\vec{x})$. We will illustrate this with the data from the EuroQol Rotterdam 1991 Survey (Essink-Bot et al., 1993; Agt et al., 1994). Before presenting our estimates, we should consider six methodological problems:

(1) *The scales*: What are the characteristics of the scales of V and the dimensions $\vec{x}$: are we dealing with ordinal, interval or ratio scales? We will prove that this is an insoluble problem and, therefore, we have to make some assumptions. We will argue however that these assumptions do not affect our main question, namely the estimation of the parametric model.

(2) *Aggregation*: Should we base our model on individual data or should we consider aggregated data? In the latter case, we must choose between various metrics, such as the mean, median or mode.

(3) *The criterion*: Which criterion should be used for the estimation of the value function? Various options are available. For instance, we may minimise distances between observed and predicted values or between observed and predicted orderings. In this paper, we limit ourselves to least squares estimations.

(4) *The model*: Which value function should be chosen? Again, there are various possibilities, but we will limit ourselves to linear models with and without interactions between the dimensions.

(5) *The health states*: Which health states should be chosen for the empirical evaluation? We will demonstrate that the choice of the model and the choice of specific health states is related.

(6) *The respondents*: Whose values should be taken into account? For instance, should we consider the values of patients or the values of the general public?

Moreover, what should we do about subjects who give 'inconsistent' values to the health states?

1.1. The scales

In the standard EQ-5D questionnaire, respondents are asked to value 13 health states plus unconsciousness using a visual analogue scale. This scale, often called *the EQ-5D Thermometer*, has *best imaginable health state* at the top with a value of 100. The bottom is labelled *worst imaginable health state* and assigned a value of 0 (Essink-Bot et al., 1993). Whatever method is used for the valuation of the health state, for example, the EQ-5D Thermometer, Standard Gamble or Time Trade-off, it is not possible to know beforehand if the response V has ratio or interval properties. However, a parametric model *requires* responses at interval level, otherwise the unrecorded values cannot be predicted. Consequently, we need to make the assumption that the response V has at least interval properties. For instance, let us assume that by scoring health states on the thermometer, respondents are able to order the various health states and that they are able to use distances on the thermometer to weigh the differences. Consequently, we assume an interval scale at *the response level*: a difference between 20 and 40 equals the difference between 40 and 60. Note that by making this assumption, nothing is said about the meaning of the values on the thermometer. Whether they represent *utility* or something else is not of concern here. We only need the assumption to estimate the unrecorded values on the scale of the thermometer using a parametric model.

In the sections above, we discussed the assumption that the response scale (in our case, the EQ-5D visual analogue scale or thermometer) has interval properties. However, assuming the same for the scales of the five health dimensions is something completely different. It is not difficult to see the limitations of this assumption, as the dimensions are constructed out of three clearly *ordinal* descriptors: 'no problems', 'some problems' and 'many problems'. If we attribute the scores 1, 2 and 3 to the three descriptors of $\vec{x}$, we assume that the distance between the worst descriptor ($x_i = 3$) and the intermediary descriptor ($x_i = 2$) is the same as the difference between the best descriptor ($x_i = 3$) and the intermediary descriptor ($x_i = 2$). One can avoid this assumption by estimating the 'true' value of the intermediary descriptor. This can be performed by introducing an additional parameter in the parametric model. However, this approach makes certain assumptions. We have already argued that we are unaware of the interval properties of the responses on the thermometer (V). This means that we cannot test the scale properties of $\vec{x}$ just on the basis of the response on V: the values of $\vec{x}$ depend on the assumptions about the interval proportions of the response V.

This identification problem is not unique for the EQ-5D, nor for any health classification system, but has been documented as a general psychometric problem (Gescheider, 1988). If a respondent perceives a stimulus, then there must be a

stimulus transformation function (f_1) that determines the relationship between the stimulus (S) and the magnitude of the sensation (ψ),

$$\psi = f_1(S). \tag{2}$$

In our case, S is the score of the health state on the five EQ-5D dimensions ($\vec{x}$) and ψ is the sensation associated with the health state, in other words: the value. However, as Shepard (1981) points out, a second transformation function is required in order to make the magnitude of the sensation (ψ) manifest as a response (R),

$$R = f_2(\psi). \tag{3}$$

In our case, this *response transformation function* (f_2) is the relationship between the value of the health state (ψ) and the mark on the thermometer (V). Therefore, the task of valuing health states is a combination of two functions: a stimulus transformation function (f_1) and a response transformation function (f_2). The observable relationship (f_3) between the stimulus (S) and the response (R), in our case $\vec{x}$ and V, is a substitution of Eq. (3) in Eq. (4):

$$R = f_3(S) = f_2[f_1(S)]. \tag{4}$$

Since the intervening variable ψ is not observable, we do not know if the observable function f_3 (in our case, $V(X)$), is determined by the scale properties of the stimuli ($\vec{x}$) or by the scale properties of the response (V). As Gescheider pointed out in 1988: ''*Unless one of the two component functions of f_3 (i.e., f_1 or f_2) is known, it is impossible to determine the other by knowledge of the experimentally determined f_3.*'' Therefore, we are ignorant of the scale properties of V and $\vec{x}$.

As Gescheider proves, we cannot know if V or $\vec{x}$ are interval scales in nature. In view of this identification problem, we assumed a linear response transformation function (f_2) in order to be able to use a parametric model to estimate the unrecorded values. By making this assumption, we limited ourselves to the determination of the stimulus transformation function (f_1), which will result in weights for x_i. Assuming such a linear transformation function often occurs in psychometric studies and is not something special for the estimation task presented in this paper. For instance, the assumption of a linear response transformation function (f_2) is a crucial assumption in the psychometric work of Stevens (Gescheider, 1988). Again, we would like to emphasise that by making this assumption nothing is revealed about the 'real interval properties' of the values V. In fact, we prove that it is impossible to test the response scale proportion by just valuing health states.

Having assumed an interval scale at the response level, we may choose to use the scores as they are or to normalise them. Using the scores as they are, means that we do not take into account that respondents may use different ranges of V. For instance, scores of respondents who valued the best health state (11111) at 100

and the worst health state (33333) at 20, will be combined with scores of respondents who gave a maximum value of 80 and a minimum value of 20. If we are interested in the relative differences between the health states, the different value ranges cause additional differences between the responses of the respondents. These extra differences can be reduced if we make the value range comparable by normalising them on the basis of the values assigned to 11111 and 33333. The normalisation that results in a value range of 0–100 is shown in Eq. (5).

$$V'(X) = \frac{V(X) - V(33333)}{V(11111) - V(33333)} \tag{5}$$

The assumption of an interval scale at the response level is still made, but responses are aggregated on the basis of the relative distance from the extreme health states instead of the distance to the extremes of the thermometer.

The values can also be normalised using states other than 11111 and 33333. For instance, in some EQ-5D investigations, values on the thermometer are normalised using the value of the state death. Another variation is to first aggregate over subjects and then normalise these means, medians, etc. Normalisation can also be performed at the end, if one normalised the predicted scores of the model. It should be noted that in the last two variations, individual difference in the range of V is not accounted for.

If subjects value an intermediate health state (for instance, 11112) higher or lower than the health state used as the standard for the normalisation (for instance, 11111 and 33333), values above 100 and below 0 will appear. These values have no theoretical upper or lower limit. Therefore, extreme normalised values could appear and dominate the predictions. In such circumstances, it may be reasonable to excluded these ‘inconsistent’ subjects from the analysis.

1.2. Aggregation

Before we can estimate a model, we have to decide whether we analyse aggregate data or analyse data on an individual level. We may choose not to aggregate the individual responses beforehand and estimate the model on the basis of the individual responses. Alternatively, we may aggregate the individual responses in advance, using one of the measures of central tendency, such as the mean, median or mode. The first procedure includes the individual error in the error term of the model, while the second procedure keeps this individual error out of the model.

When the model is estimated on aggregate data, a choice can be made between various metrics of central tendency. Using averages might be regarded as a pure utilitarian procedure, while using the median implies the ‘median voter model’ based on public choice theory (Williams, 1993, p. 299). In fact, by choosing a

measure of central tendency, we are making choices about whose values should count (Williams, 1992, p. 10), which is a subject of debate. For instance, by using the median, all responses have an equal influence on the central value. When using the mean however, extreme responses are given extra weight.

If using aggregated data, the degrees of freedom to fit the model are determined by the number of empirically valued health states. If individual data is used, the degrees of freedom are largely determined by the number of subjects. Because the number of subjects is usually much higher than the number of empirically valued health states, it seems obvious to use individual data in order to increase the power to detect significant parameters. However, as each subject can value only a limited number of health states, subjects are often divided into groups, each valuing different health states. In this instance, subjects do not value the same health states, which may introduce additional group effects. If the nested structure is ignored, the standard errors of the estimates will be underestimated and so coefficients will be judged to be significant when they are not. Thus, the usual *fixed effect* models may not be well equipped for the situation where data is clustered. More sophisticated models such as the *multilevel* or *random effect* models may be more appropriate and have now been introduced in this field of science (Goldstein, 1995; Dolan, 1997).

1.3. The criterion

If the value function $V(\vec{x})$ is estimated, a criterion function is required to describe the fit of this model. Various criteria are available. A natural option is to minimise the differences between the observed and predicted valuations using various metrics. In this paper, we considered only variations of least squares estimations. The main reason for this choice was its feasibility and the fact that the results can easily be interpreted. We should, however, realise that there are several alternatives. One alternative might be the minimisation of the differences between the observed and predicted ranks. We can also perform a log- or similar transformation before we estimate the model and then minimise the differences between the observed and predicted transformed scores. Such a transformation may sometimes normalise the distribution of the error term but caution is needed as in the case of a log transformation: the appropriate back transformation is not simply the exponential function of the transformed scores. If an exponential function is used, the converted scores will be neither unbiased nor consistent and, therefore, another back transformations should be considered (Rutten-Mölken et al., 1994).

If we use ordinary least-squares, it is important to realise that the 'common' interpretation of the size R^2 is not possible. The common interpretation (0.20 is 'bad', 0.90 is 'good') is only valid when the 'observations' have been sampled randomly. The subjects who value the health states may have been sampled randomly, but it should be noted that these subjects are not the units of observation. The units of observation are the health states, and the health states are chosen

by the investigators. Because investigators will usually choose health states that cover the whole range of possible health states equally, the distribution of the values is not normal but flat, which results in high R^2. In Section 3, we demonstrate the dependency of R^2 on the choice of the health states. Of course within a given selection of health states, a relatively higher R^2 still represents a relatively better fit of the model. The absolute size however, should be interpreted with care.

1.4. The model

Some of the components of the model are already determined by the choices we made earlier. We have assumed that each health state can be characterised by a five-dimensional vector of scores $\vec{x}$, and that there exists a linear relationship between $\vec{x}$ and the value V given on the thermometer. This linear relationship can be described by a value function $V(\vec{x})$ about which we have to make two assumptions. First, we assumed the value function is continuous and twice differentiable in its arguments: an infinitely small change in any dimension leads to an infinitely small change in the value attributed to the health state. This means that the *latent constructs* that are represented by $\vec{x}$ and $V(\vec{x})$ are continuous in nature. In other words, the *underlying trait* of both $\vec{x}$ and $V(\vec{x})$ does not make any 'jumps', despite the fact that $\vec{x}$ is discrete. Second, we assumed that the first order derivatives are positive, in other words, a better score on each dimension leads to a higher valuation of the corresponding health state. There are a number of different functional forms of the value function that we can consider. For practical reasons, we considered only three additive forms of the value function. An additive form means that the contributions of the different dimensions can be summed to produce the value of the health state. The first form assumes not only a linear response transformation function, but also a linear stimulus transformation function. In other words, in this model, we not only assume that the scale V holds interval proportions at the response level, but that the same is true for the scores on the stimuli, namely, 1, 2 and 3 on x_i. Thus, we assume that the intermediary score 2 is the mid point of the scales x_i.

$$V(X) = \alpha + \sum_{i=1}^{5} \omega_i x_i + e \tag{6}$$

In Eq. (6), ω_i represents the weight of the different dimensions. For instance, if the first dimension (mobility) plays a more important role than the second dimension (self-care), then the weight of the first dimension (ω_1) should be higher than the weight of the second dimension (ω_2). α is a constant and e is an error term. As both V and x_i are assumed to be interval scales, Eq. (6) is simply a multiple regression in which V is the dependent variable, x_i are the explanatory (independent) variables and α is the intercept. In the multiple regression, the

regression coefficient B_i represents ω_i and can be used as an estimate for these weights.

If we weaken the assumption of a linear stimulus transformation function, we have to estimate new values for x_i. As there are only three possible scores per dimension, the nonlinear relationship can only be expressed by the intermediary descriptor of x_i, in our case, the score two. We can describe this by adding five constants m_i to Eq. (6), if x_i takes the value two. The values of the constants m_i can then be estimated by adding five dummy variables in the regression which take the value one if $x_i = 2$ and zero otherwise.

$$V(X) = \alpha + \sum_{i=1}^{5} \omega_i x_i + \sum_{i=1}^{5} m_i + e \tag{7}$$

On the basis of the value of m_i a new value for the intermediary descriptor of x_i can be calculated.

$$x_i' = x_i + \frac{m_i}{\omega_i} \tag{8}$$

Instead of Eq. (7), one can also use more conventional dummies: one dummy per dimension that indicates that $x_i \geq 2$, and one dummy that indicates that $x_i \geq 3$. In principle, the results will be the same. We have chosen Eq. (7), because this equation more clearly demonstrates the relaxation of the linear stimulus transformation assumption.

Eq. (6) does not take interactions between dimensions into account. Eq. (9) is an extension of Eq. (6) with a first order interaction term. In this equation, the values of the levels x_i are assumed to be known beforehand. If estimating ω_{ij}, ω_i and x_i simultaneously, m_i should be included in Eq. (9). This new equation will be complex, as the m_i term will be different for all different interactions. To our knowledge, this model is not yet described nor used, although sometimes investigators test the interaction terms $x_i x_j$ separately in Eq. (7) (Dolan, 1997).

$$V(X) = \alpha + \sum_{i=1}^{5} \omega_i x_i + \sum_{i=1}^{5} \sum_{j=2}^{5} \omega_{ij} x_i x_j + e \qquad (i < j) \tag{9}$$

The Health Utility Index of Torrance et al. (1996) is based on a model comparable to Eq. (9), although in their model the interactions are taken into account in a more restricted way (Eq. (10)). As one can see in Eq. (10), the interactions between the dimensions do not depend on the specific combination of the levels: *mutual utility independence*. These assumptions about the interactions are not directly based on empirical data, but derived from a theoretical model: *multiattribute utility theory* (MAUT) as described, for example, by Keeney and Raiffa (1976). As the assumptions about the interactions simplify the model, Torrance et al. needed only a minimal number of empirically valued health states

in order to estimate this multiplicative model. The value function can be estimated with nonlinear regression analysis.

$$V(X) = \frac{1}{\alpha}\left[\prod_{i=1}^{5}(1 + \alpha\omega_i x_i) - 1\right] + e \tag{10}$$

Torrance et al. scaled both V and $\vec{x}$ on a [0...1] scale, instead of a [0...100] scale and a [1...3] scale. An advantage of such a transformation is that the parameters α and ω_i now have a generally accepted interpretation, as formulated in Eq. (11) (Keeney and Raiffa, 1976, p. 240; Torrance et al., 1996):

$$\begin{aligned}
&\text{If } \sum_{i=1}^{5}\omega_i > 1, \text{ then } -1 < \alpha < 0\\
&\text{If } \sum_{i=1}^{5}\omega_i = 1, \text{ then } \alpha = 0\\
&\text{If } \sum_{i=1}^{5}\omega_i < 1, \text{ then } \alpha > 0
\end{aligned} \tag{11}$$

If $\alpha = 0$, the dimensions of health assume *additive independence*. This means that a step from 1 to 2 in one dimension would always have the same influence, irrespective of the levels of other dimensions. If $-1 < \alpha < 0$, and mutual utility independence exists, the dimensions of health are said to be *substitutes*: ''*...an improvement in one* (dimension of health) *is relatively satisfying, while an improvement on two or more* (dimensions of health) *is not that much better.*'' (Torrance et al., 1982, p. 1049). If $\alpha > 0$, then the dimensions of health are said to be *complements*: ''*...an improvement on any one* (dimension of health) *is not very useful, while a simultaneous improvement on several* (dimensions of health) *is much better*.''

If we choose to normalise the values on the thermometer, it seems reasonable to give the predicted values the same range by assuming that $V'(33333) = 0$ and $V'(11111) = 100$. This means that the response range restricts the weights ω_i: the differences between the best (11111) and the worse health state (33333) must then be 100:

$$\begin{aligned}
&V(11111) - V(33333) = 100 \Leftrightarrow\\
&\alpha + \omega_1 1 + \omega_2 1 + \omega_3 1 + \omega_4 1 + \omega_5 1 - \alpha - \omega_1 3 - \omega_2 3 - \omega_3 3 - \omega_4 3 - \omega_5 3\\
&\quad = 100 \Leftrightarrow\\
&\omega_1(1-3) + \omega_2(1-3) + \omega_3(1-3) + \omega_4(1-3) + \omega_5(1-3)\\
&\quad = 100 \Leftrightarrow\\
&-2(\omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5) = 100 \Leftrightarrow\\
&\omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5 = -50
\end{aligned} \tag{12}$$

Besides restricting the range to [0,100], we have to force one endpoint to its desired value, for instance $V'(33333) = 0$. If we substitute this value in the last line of Eq. (12), we can calculate the value of the intercept α:

$$
\begin{aligned}
V'(33333) = 0 \ \wedge \ & \omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5 \\
& = -50 \Leftrightarrow \\
0 = \alpha + \omega_1 3 + \omega_2 3 + \omega_3 3 + \omega_4 3 + \omega_5 3 \ \wedge \ & \omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5 \\
& = -50 \Leftrightarrow \\
0 = \alpha + 3(\omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5) \ \wedge \ & \omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5 \\
& = -50 \Leftrightarrow \\
0 = \alpha - 150 & \Leftrightarrow \\
\alpha = 150 & \qquad (13)
\end{aligned}
$$

If we substitute this value of α in Eq. (6), while at the same time we know from Eq. (12) that $\omega_5 = -50 - \omega_1 - \omega_2 - \omega_3 - \omega_4$, we get:

$$
\begin{aligned}
V'(X) = \alpha + \sum_{i=1}^{5} \omega_i(x_i) + e \ \wedge \ \alpha = 150 \ \wedge \ & \\
\omega_5 = -50 - \omega_1 - \omega_2 - \omega_3 - \omega_4 & \Leftrightarrow \\
V'(X) = 150 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + \omega_4 x_4 & \\
+ (-50 - \omega_1 - \omega_2 - \omega_3 - \omega_4) x_5 + e & \Leftrightarrow \\
V'(X) = 150 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + \omega_4 x_4 - 50 x_5 - \omega_1 x_5 - \omega_2 x_5 & \\
- \omega_3 x_5 - \omega_4 x_5 + e & \Leftrightarrow \\
V'(X) + 50 x_5 - 150 = \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + \omega_4 x_4 - \omega_1 x_5 - \omega_2 x_5 & \\
- \omega_3 x_5 - \omega_4 x_5 + e & \Leftrightarrow \\
V'(X) + 50 x_5 - 150 = \omega_1(x_1 - x_5) + \omega_2(x_2 - x_5) + \omega_3(x_3 - x_5) & \\
+ \omega_4(x_4 - x_5) + e. & \qquad (14)
\end{aligned}
$$

We can now estimate the weight ω_i by a multiple regression in which the intercept is set at zero; $(x_1 - x_5)$, $(x_2 - x_5)$, $(x_3 - x_5)$ and $(x_4 - x_5)$ are the new independent variables and $(V' + 50 x_5 - 150)$ is the new dependent variable. Afterwards, we can calculate ω_5 and the standard error from Eq. (12).

Unfortunately, it has not been possible to find an equation that would do the same for a model that contains interactions between the dimensions. All models we considered resulted in a predicted value range of less then 100. Therefore, there is not yet a good parametric model that predicts normalised values and allows for interactions between the dimensions. Forcing the predicted values to range from 100 to 0 must therefore be carried out on an ad hoc basis, for instance, by normalising the predicted values afterwards.

1.5. The health states

After all choices have been made, we are still not ready to estimate $V(\vec{x})$. First, we must select a subset of health states from the 243 possible EuroQol health states. The choice of the health states relates to choice of the model and to the method the investigator uses to elicit health states from the subjects. Generally speaking, one can distinguish two such methods: *statistically inferred strategies* and *explicitly decomposed strategies* (Froberg and Kane, 1989a).

The EuroQol Group and the group of Kaplan et al. (Kaplan et al., 1976) use the *statistically inferred* method. The investigator asks the subject to value a number of (complete) health states. The weights of the different dimensions, their levels and possible interactions are all deduced by decomposition afterwards, with the use of a model. Because one is unaware of these values, a model is often chosen that can estimate these parameters, such as the models represented in Eqs. (7) and (9). If we want to estimate these parameters, we have to choose health states that represent the whole valuation space as closely as possible. That means that one should include bad, good and intermediate health states in the valuation sample. Furthermore, in order to estimate the interactions, a selection of health states should be made that maximises the number of combinations of the levels across the dimensions.

Torrance and colleges use the *explicitly decomposed* method: the weights of the different dimensions, their levels and their possible interactions are all determined separately. The weights for the different dimensions are determined by valuing the so called *corner states*: health states that hold all but one dimension at the best level, and that one dimension is set at the worst level. In terms of the EQ-5D these states would be 31111, 13111, 11311, 11131 and 11113. In the studies of Torrance, the subjects valued these corner states using a visual analogue scale with the best possible state at the top (11111) and the worse state at the bottom (33333). The values of the levels within the dimensions are determined holding the levels of the other dimensions at a fixed level: the subjects were asked to assume that 'all other aspects of your health and abilities are normal'. Subjects valued the intermediate levels (21111) again with a visual analogue scale that now had the best level at the top (11111) and the matching corner state at the bottom (31111). The explicitly decomposed method is most often used in combination with the model based on mutual utility independence (Eq. (10)). By making these explicit assumptions about the interaction of the dimensions, the values of all possible health states can be estimated using the empirical valuations mentioned above. Using the explicitly decomposed method, the corner states are chosen for the valuation task. It could be said that the explicitly decomposed method estimates the valuation space from the outside. On the other hand, the statistically inferred method uses *interior states* (Hakim and Pathak, 1995) and extrapolates these values towards the borders of the valuation space. Using corner states has the advantage that there are only a small number of them. For instance, using the

explicitly decomposed method for the valuation of the EuroQol health states, would require only 10 values: the corner states 31111, 13111, 11311, 11131, 11113 for weighting the dimensions and 21111, 12111, 11211, 11121, 11112 for weighting the levels. Of course, it is only possible to use so few health states, by making strong assumptions, namely mutual utility independence.

In choosing health states, one is limited by the fact that some health states are difficult to imagine, for instance 'confined to bed' but with 'no problems' on the other dimensions (31111). This problem is especially encountered using the explicitly decomposed methods, because these depend heavily on the valuations of such extreme health states. In their earlier work, Torrance et al. (1996) tried to get around this problem, by estimating the value of the corner state 31111 with more probable states, such as 33111 (a 'double' corner state), or 21111 (a 'backed-off' corner state). However, this increases the complexity of the model and also increases the influence of inconsistent responses on the estimations. Therefore, in later work, they adapted their health classification system in an attempt to make the dimensions less *structurally* dependent of each other (Feeny et al., 1995, p. 495). Structural independence between the dimensions did not play an important role when the EuroQol Group chose their dimensions (The EuroQol Group, 1990; Williams, 1995). Indeed the research of the EuroQol Group is concentrated on statistically inferred methods, avoiding valuations of corner states.

1.6. The respondents

The last decision concerns which respondents to use in the evaluation. In other words: whose values are taken into account? There has been much debate about the differences between the values of patients and non-patients (Torrance, 1986: Froberg and Kane, 1989b; Carr-Hill, 1991; Hadorn, 1991; Williams, 1991; Selaï and Rosser, 1995). In this investigation, we limited ourselves to values given by the general public, representing the so called *societal viewpoint* (Hadorn, 1991; Gold et al., 1996).

Having made this choice, we also have to decide how to deal with missing values and 'clearly irrational responses', e.g., respondents who give a very bad health state a high value, higher than the best health state? These kinds of outliers have a large influence on mean values, especially in the case of normalised values. Furthermore, Gold et al. (1996) argue that the vales for health states should be given by ''*[. . .] a well-informed, cognitive robust, unbiased community sample.*'' (p. 106). In an earlier stage of this investigation, van Hout and McDonnell (1992) adopted this point of view rather rigorously, and dropped all respondents who showed signs of irrationality or misinterpretation. A similar protocol was followed by Torrance et al., 1996. On the other hand, many social scientists will argue that signs of irrationality and misinterpretation and outliers are empirical findings and should be incorporated into the data set. In the present study, we adopted an intermediate viewpoint, and used only a few exclusion criteria for the 'original values'. This means that small mistakes are not seen as signs of misinterpretation,

but as measurement error. We used a more restricted selection of the subjects for the analyses of the 'normalised values', because outliers can easily dominate these values.

2. Method

In this paper, we estimated the value function $V(\vec{x})$ on the basis of data from the EuroQol Rotterdam 1991 Survey. This general population survey is already described in detail in the paper of Essink-Bot et al. (1993) and Agt et al. (1994). Each respondent valued 16 health states using the EQ-5D thermometer. By using different versions of the questionnaire, we obtained values for 25 health states. These versions were sent to 1400 households in Rotterdam. We selected only questionnaires with two or less missing values. This is also the selection used by Essink-Bot et al. (1993). They made this selection under the assumption that *''...if only one or two states were missing the respondent had essentially understood the task.''* We used a more restricted selection of the subjects for the analyses of the normalised values. We disregarded the responses of respondents who valued intermediate states higher than the best health state (11111) and/or lower than the worst (33333). Furthermore, we disregarded respondents who valued the best and the worst states equally.

We estimated the parameters of the model on ordinary least squares and, if possible, on multilevel analysis. This was only possible with the least complicated models. For instance, we were not able to find a satisfying equation that would enable us to estimate normalised values using multilevel analysis.

We estimated the parameters of the multiplicative model on the basis of Eq. (10) and the normalised average values of V and the assumption that $x_i = 2$. Instead of using the explicitly decomposed strategy, as Torrance et al. (1996) proposed, we used the statistical inferred strategy in combination with this multiplicative model. In order to facilitate the interpretation of α and ω, in the case of Eq. (10), we transformed the values on V to a $[0 \ldots 1]$ scale and the scores on the dimensions of $\vec{x}$ from {1,2,3} to {1,0.5,0}.

3. Results

The respondents returned 980 questionnaires (70%), of which 643 questionnaires (46%) had two or less missing values. Essink-Bot et al. (1993) present the background variable of this response and described an in-depth non-response investigation. They reported that the similarities between responders and non-responders in terms of background variables were more striking than the differences. Agt et al. (1994) measured the test–retest reliability of this data, and describe this reliability as good.

Table 2 presents the mean values of the response. Because we analysed the same data, these mean values are the same as in Table 2 of Essink-Bot et al. (1993). There were 135 (21%) subjects who had valued at least one health state below 33333 and 73 (11%) subjects who had valued at least one health state above 11111. Some of these responses caused extreme normalised values: 18 subjects had normalised scores with a minimum value ranging from −10 to −4750; 22 subjects had normalised scores with a maximum value ranging from 100 to 1200. These outliers were excluded from the analyses of the normalised values, using the criteria mentioned in Section 2.

Table 3 presents estimates of the linear models without interaction terms (Eqs. (6) and (7)). The estimates in the first columns are based on the mean and the

Table 2
Empirical values of the health states for respondents with two or less missing values ($N = 643$)

States	Original values					Normalised values [0…100]				
	Mean	SD	Median	Mode	*N*	Mean	SD	Median	Mode	*N*
11111b	92.31	13.16	97	100	639	100.00	0.00	100	100	389
11111a	92.25	13.63	97	100	639	100.00	0.00	100	100	389
11211	80.47	14.38	80	80	331	82.28	13.90	85	80	197
11121	73.58	18.48	75	70	332	73.70	19.13	75	80	197
11112	73.44	18.69	75	80	341	72.74	19.39	77	80	219
12111	67.93	23.68	70	80	337	70.36	20.67	73	80	219
21111	62.93	23.18	68	70	333	62.84	20.67	66	50	219
11221	65.48	18.09	65	60	300	66.21	17.66	68	70	170
11122	59.95	20.65	60	60	333	58.55	18.72	59	60	197
21211	52.90	23.51	58	60	306	54.01	21.85	58	50	192
12212	52.71	20.13	54	50	297	52.51	18.95	52	50	170
21212	48.50	20.25	50	50	300	46.78	19.83	47	50	170
32211	45.17	23.33	45	40	338	38.56	19.44	39	0	219
21232	35.11	23.87	30	20	333	27.57	17.49	28	20	197
23223	29.89	22.56	30	30	308	21.86	15.84	21	0	192
22233	27.06	23.15	20	20	333	16.73	13.48	14	0	197
33321	26.31	23.01	20	20	332	16.12	13.41	13	0	197
22323	25.96	22.98	20	20	339	17.03	12.35	16	0	219
32233	24.87	23.44	20	0	307	14.95	14.51	11	0	192
22333	24.81	22.71	20	10	306	16.27	14.68	13	0	192
23332	21.24	21.31	15	0	306	12.79	12.40	11	0	192
32333	20.79	22.65	15	10	297	10.54	11.44	8	0	170
33332	20.65	21.93	15	20	308	9.85	9.56	8	0	192
33233	19.80	21.46	15	10	296	10.71	12.41	9	0	170
23333	15.67	20.80	10	0	299	7.36	9.98	5	0	170
33333b	14.39	23.20	5	0	642	0.00	0.00	0	0	389
33333a	13.33	23.08	5	0	642	0.00	0.00	0	0	389

States 11111 and 33333 were presented twice; the first presentation (a) is used in the calculations. The values are sorted according to the median of the original values. The original values are taken from the investigation of Essink-Bot et al. (1993).

Table 3
Parameter estimates without interaction

	Original values								Normalised values					
	Mean		Median		Individual OLS		Individual MLA		Mean		Median		Individual OLS	
	x_i	x'_i	x_i	x'_i	x_i	x'_i	x_i	x'_i	x_i	x'_i	x_i	x'_i	x_i	x'_i
α	111.12	116.45	119.96	125.73	113.731	119.23	114.00	119.30	150.00	150.00	150.00	150.00	150.00	150.00
ω_1	−9.17	−10.88	−9.23	−10.98	−9.509	−11.35	−9.521	−11.52	−13.29	−17.41	−17.41	−16.87	−14.06	−17.83
ω_2	−5.07	−7.16	−5.48	−7.44	−4.775	−7.27	−4.757	−7.41	−1.58	−8.51	−8.51	−9.14	−1.62	−8.75
ω_3	−7.72	−4.73	−9.68	−6.83	−8.169	−5.00	−8.302	−4.93	−16.81	−8.75	−8.75	−8.09	−16.17	−8.12
ω_4	−4.79	−5.97	−7.43	−8.65	−5.049	−6.00	−4.992	−5.79	−7.68	−7.42	−7.42	−8.62	−7.62	−7.43
ω_5	−7.28	−4.12	−7.51	−4.34	−7.454	−4.39	−7.528	−4.41	−10.64	−7.91	−7.91	−7.28	−10.53	−7.86
x_1	2.00	2.78	2.00	2.68	2.00	2.73	2.00	2.75	2.00	2.44	2.44	2.47	2.00	2.44
x_2	2.00	2.63	2.00	2.73	2.00	2.62	2.00	2.63	2.00	2.56	2.56	2.72	2.00	2.75
x_3	2.00	2.33	2.00	2.23	2.00	2.36	2.00	2.33	2.00	2.45	2.45	2.44	2.00	2.44
x_4	2.00	2.72	2.00	2.56	2.00	2.82	2.00	2.81	2.00	3.34	3.34	3.13	2.00	3.31
x_5	2.00	3.41	2.00	3.48	2.00	3.45	2.00	3.44	2.00	3.30	3.30	3.37	2.00	3.27
R^2 (adj.)	0.93	0.97	0.94	0.97	0.54	0.56			0.75	0.92	0.92	0.95	0.63	0.75

The parameter can be used in a linear model that estimated the values of the EQ-5D health states. α = a constant; ω_1 = the weight of the EQ-5D dimension mobility; ω_2 = item for self-care, etc.; x_1 is the value of the in-between descriptor of mobility; x_2 item for self-care, etc.; x_i is the in-between descriptor without rescaling. The value of x_i is therefore always 2. x'_i is the estimated value of the in-between descriptor. The predicted values based on the normalised values have a range of 0–100. Individual OLS = parameters based on individual values using ordinary least squares; MLA = multilevel analysis.

median without normalisation. Estimates can be made based on the mode using the data presented in Table 2. Table 3 also presents the adjusted R^2 of the different models. It should be noted that the common interpretation of this parameter is inappropriate here, because the size of this parameter depends on the distribution of the health states. For instance, if we fit the most simple model (Eq. (6)) on the individual data only using the states 21111, 12111, 11211, 11121 and 11112 the R^2 would be as low as 0.078. If we then replace 11112 by 33333, the R^2 increases to 0.628. This is even higher than the R^2 for all 25 states as presented in Table 3. In multilevel analysis there are several 'R^2 like measures', which makes the interpretation different from the models based ordinary least squares.

Table 4 presents the estimates of the interactions between the dimensions. Because we were unable to find an equation that would estimate both the interaction terms and force the predicted values to the range from 0 to 100, the predictions of the normalised values do not cover the full range between 100 and 0. For instance the value of state 11111 is 93.19 and the value of 33333 is 4.38 (if $x_i = 2$).

Table 4
Parameter estimates with interaction parameters. Estimates based on the individual values. ω_{12} = the first order between mobility and self-care

	Original values		Original values		Normalised values	
	Ordinary least squares		Multilevel analysis		Ordinary least squares	
	x_i	p	x_i	p	x_i	p
α	165.19	0.00	165.30	0.00	191.41	0.00
ω_1	−29.07	0.00	−29.14	0.00	−33.38	0.00
ω_2	−18.85	0.00	−18.80	0.00	−20.71	0.00
ω_3	−9.14	0.00	−9.26	0.00	−9.91	0.00
ω_4	−20.07	0.00	−20.03	0.00	−29.42	0.00
ω_5	−17.16	0.00	−17.20	0.00	−22.74	0.00
x_1	2.00		2.00		2.00	
x_2	2.00		2.00		2.00	
x_3	2.00		2.00		2.00	
x_4	2.00		2.00		2.00	
x_5	2.00		2.00		2.00	
ω_{12}	9.08	0.00	8.90	0.00	8.70	0.00
ω_{13}	−3.17	0.01	−3.06	0.00	−3.64	0.00
ω_{14}	0.39	0.71	0.32	0.32	2.53	0.01
ω_{15}	4.10	0.00	4.27	0.00	3.03	0.01
ω_{23}	1.15	0.36	1.27	0.07	0.18	0.87
ω_{24}	−2.94	0.00	−2.97	0.00	−1.24	0.16
ω_{25}	−0.82	0.35	−0.77	0.06	−0.80	0.30
ω_{34}	4.63	0.00	4.67	0.00	4.30	0.00
ω_{35}	−0.76	0.50	−0.94	0.23	0.48	0.62
ω_{45}	3.41	0.00	3.41	0.00	4.39	0.00
R^2 (adj.)	0.56				0.84	

Table 5
Parameter estimates for the multiplicative model (Eq. (10)) on the basis of normalised average values

Parameters	Estimates	Asymptotic SE
α	3.612	1.283
ω_1	0.143	0.030
ω_2	0.081	0.023
ω_3	0.062	0.025
ω_4	0.084	0.022
ω_5	0.096	0.025
R^2	0.97	

V and X are both transformed to a $[0 \ldots 1]$ scale.

The parameters' estimates of the multiplicative model are presented in Table 5. As $\Sigma\omega_i < 0$ and $\alpha > 0$, all dimensions should be interpreted as complements. As indicated previously, this means that an improvement on any one of the dimensions is not very beneficial, while a simultaneous improvement on several is much better. Note that again the estimated values of this model with interactions are not normalised: the estimated value of 11111 is 0.889, lower than 1.00.

4. Discussion

We used data from a postal EQ-5D survey to estimate several value-functions. These value functions can be used to assign values to health states that were not recorded. Before presenting our estimates, we distinguished six topics that are directly related to our main question: the characteristics of the scales, the choice of the aggregation mode, the choices of a criterion function, the specification of the model and choices of the health states and respondents. Various choices had to be made in relation to these topics. An important choice was to interpret values as scores on an interval scale. We also chose to minimise the ordinary least-squares and we based our estimates on the means, the medians and on all individual data. After these choices, we considered several linear and nonlinear models. For example, we took into account that the intermediary descriptors on each dimension might not be equally spaced and estimated their corresponding values. Furthermore, we looked at the way interaction could be incorporated in the model. Finally, we considered the choice of health states and the respondents.

It appears that all models fit the data rather well in terms of R^2. However, as we have explained before, the R^2 is difficult to interpret and we suspect that all models are subject to mis-specification. An example of such mis-specification is that the estimated value for the intermediary descriptor of anxiety/depression is sometimes above the value of the worst descriptor (Table 3). On the basis of the present data, it is difficult to see what causes this. It could be a 'real' phenomenon, but it might also be an artefact caused by the presentation of the health states. In another investigation we found that the place of the health states on the

pages influences the value. If the health state is placed at the top of the page, the values will be relatively high; when the box is place at the bottom, the value is relatively low. In other words, subjects minimise the length of the line they have to draw. Furthermore, mis-specification disappeared when the 24 health states were presented randomly (Busschbach et al., 1997).

The differences between the parameters based on the raw individual responses and the parameters based on the mean values of the health states (Table 3) were minimal. This is because we chose to minimise the sum of squared residuals: in both cases the mean values were used for estimating the parameters. Small differences occurred because the models based on individual data take into account differences in sample sizes on which the values of the health states are based (see '*N*' in Table 1). Larger differences may emerge if we use normalised data, because groups of health states are then associated with the values of their accompanying health states 11111 and 33333. This is not clearly seen in Table 3, because the estimates for the individual data were based on more consistent subjects than the estimates on average data. If we had estimated them on responses from subjects with 2 or less missing values (as we did with the average data), large differences would occur.

The difference between the ordinary least square models and the models based on multilevel analysis were small. In fact, it is tempting to conclude that it does not make a differences, and one can rely on the well known fixed effect models. This is probably caused by the relatively high number of health states in which the respondents are nested. Indeed, Goldstein indicated that if there are a relatively high number of units of analysis on level 2 (in our case, the health states) as compared to level 1 (the respondents), the precision of an ordinary regression analysis will be improved considerably (Goldstein, 1995, p. 2). Differences might be larger in other samples with a lower number of health states. On the other hand, although from a theoretical point of view the multilevel analysis might be more appropriate than fixed effect models, differences might be small when one chooses a design incorporating many health states.

We used the statistical inferred strategy in combination with this multiplicative model. The relatively good fit in terms of R^2, shows the use of this multiplicative model is not conditional on the use of explicitly decomposed strategies. However, note that the explicitly decomposed method can only be used in combination with multiplicative models.

The parameters' estimates of the multiplicative model indicated that all dimensions should be interpreted as complements. The interpretation of 'complements' was done in terms of 'getting better'. If one would like to make an interpretation in terms of getting sick, the dimensions should be interpreted as substitutes: getting worse in one dimension is already big problem, while getting worse in two dimensions is not that much worse.

From the present data set, it is not possible to test empirically many of the different choices that have to be made before the accuracy of a model can be

estimated. Currently, we are collecting a new data set that hopefully can be used for this purpose. A group of 100 students valued *all* 243 health states empirically. Using this data set, we hope to test empirically how accurate the different models are in predicting all 243 values from a specific subset (Busschbach et al., 1997).

Acknowledgements

The Netherlands' Health Research Promotion Programme and The Netherlands' Institute for Health Sciences (NIHES) funded this research project. The authors are also grateful to the Merck Foundation for financial support. Rosalind Rabin edited the English.

References

Agt, H.M.Ev., Essink-Bot, M.L., Krabbe, P.F.M., Bonsel, G.J., 1994. Test–retest reliability of health state valuations collected with the EuroQol questionnaire. Social Sciences and Medicine 39, 1537–1544.

Brooks, R., 1996. EuroQol: the current state of play. Health Policy 37, 53–72.

Busschbach, J.J.V., McDonnell, J., Hout, B.Av., 1997. Testing different parametric relations between the EuroQol health description and health valuation in students. In: Nord, E. (Ed.), Conference Proceedings of the EuroQol Plenary Meeting Oslo, October 17–19, 1996. Working Paper No. 2/97, National Institute of Public Health, Oslo.

Carr-Hill, R.A., 1991. Allocating resources to health care: is the QALY (Quality Adjusted Life Year) a technical solution to a political problem?. International Journal of Health Services 21, 351–363.

Dolan, P., 1997. Modeling valuations for the EuroQol health states. Medical Care 35, 1095–1108.

Essink-Bot, M.L., Stouthard, M.E.A., Bonsel, G.J., 1993. Generalizability of valuations on health states collected with the EuroQol questionnaire. Health Economics 2, 237–246.

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

Feeny, D., Furlong, W., Boyle, M., Torrance, G.W., 1995. Multiattribute health status classification systems. Health Utilities Index. PharmacoEconomics 7, 490–502.

Froberg, D.G., Kane, R.L., 1989a. Methodology for measuring health-state preferences: I. Measurement strategies. Journal of Clinical Epidemiology 42, 345–354.

Froberg, D.G., Kane, R.L., 1989b. Methodology for measuring health-state preferences: III. Population and context effects. Journal of Clinical Epidemiology 42, 585–592.

Gescheider, G.A., 1988. Psychological scaling. Annual Review of Psychology 39, 169–200.

Gold, M.R., Patrick, D.L., Torrance, G.W., Fryback, D.G., Hardon, D.C., Kamlet, M.S., Daniels, N., Weinstein, M.C., 1996. Indentifying and valuing outcomes. In: Gold, M.R., Siegle, J.E., Russel, L.B., Weinstein, M.C. (Eds.), Cost-Effectiveness in Health and Medicine. Oxford Univ. Press, New York, ISBN: 0-19-510824-8.

Goldstein, H., 1995, Multilevel Statistical Models, 2nd edn. Edward Arnold London, ISBN: 0-340-59529-9.

Hadorn, D.C., 1991. The role of public values in setting health care priorities. Social Science and Medicine 32, 773–781.

Hakim, Z., Pathak, S., 1995. Modelling the EuroQol data: a comparison of discrete choice conjoint and conditional utility modelling. Paper Presented at the 12th EuroQol Plenary Meeting, Barcelona, 3/10/1995.

Kaplan, R.M., Bush, J.W., Berry, C.C., 1976. Health status: types of validity and the index of well-being. Health Services Research 11, 478–507.

Keeney, R.L., Raiffa, H., 1976. Decisions with Multiple Objectives: Preferences and Value Tradeoffs. Wiley, New York, ISBN: 0-471-46510-0.

Rutten-Mölken, M.P.M.Hv., Doorslaer, E.K.A., Vliet, R.C.J.A., 1994. Statistical analysis of cost outcomes in a randomized controlled clinical trail. Health Economics 3, 333–345.

Selaï, C., Rosser, R., 1995. Eliciting EuroQol descriptive data and utility scale values from patients. A feasibility study. PharmacoEconomics 8, 147–158.

Shepard, R.N., 1981. Psychological relations and psychophysical measurement: on the status of 'direct' psychological measurement. Journal of Mathematical Psychology 24, 21–57.

Torrance, G.W., 1986. Measurement of health state utilities for economic appraisal. Journal of Health Economics 5, 1–30.

Torrance, G.W., Boyle, M.H., Horwood, S.P., 1982. Application of multiattribute utility theory to measure social preference for health states. Operations Research 30, 1043–1069.

Torrance, G.W., Feeny, D.H., Furlong, W.J., Barr, R.D., Zhang, Y., Wang, Q., 1996. Multiattribute utility function for a comprehensive health status classification system. Health Utility Index Mark 2. Medical Care 34, 702–722.

Hout, B.Av., McDonnell, J., 1992, Estimating a parametric relation between health description and health valuation using the EuroQol instrument. In: Björk, S. (Eds.), Discussion Paper No. 1. EuroQol Conference Proceedings, Lund, October 1991. IHE Working Paper 1992: 2, ISSN: 1100–4657.

Williams, A., 1991. Is the QALY a technical solution to a political problem? Of course not!. International Journal of Health Services 21, 365–369.

Williams, A., 1992. Cost-effectiveness analysis: is it ethical?. Journal of Medical Ethics 18, 7–11.

Williams, A., 1993. Priorities and research strategy in health economics for the 1990s. Health Economics 2, 295–302.

Williams, A., 1995, The measurement and valuation of health: a chronicle. Discussion Paper 136. Centre for Health Economics, University of York.