PREFERENCE-BASED INDEX MEASUREMENT OF HEALTH-RELATED QUALITY OF LIFE: WHEN DOES IT REFLECT ONLY ARBITRARY SETTINGS OF THE RESEARCHER?

UWE KONERDING*

Trimberg Research Academy, University of Bamberg, Bamberg, Germany

SUMMARY
In health economics, health-related quality of life (HRQoL) is often assessed by means of preference-based index measurement instruments (e.g. EQ-5D, SF-6D, HUI). Each instrument of this kind consists of (1) a multi-attribute classification system for distinguishing health states and (2) a scoring function which assigns a valuation to each health state distinguished within the classification system. Scoring functions are often produced according to the so-called statistical approach which consists of two steps: (1) the valuations of some of the health states of the classification system are empirically determined and (2) the scoring function values for all health states are estimated from the empirically determined valuations using a theoretical model, i.e. an index model. This approach can run into problems because the empirically determined valuations necessarily contain arbitrary settings. This article is concerned with how these arbitrary settings together with the index model affect the final scoring function values. It is shown that for many conceivable index models the final scoring function values have no empirical meaning. Only additive models with a free additive constant are appropriate for representing the information contained in the empirically determined valuation. Only these models should be used within the statistical approach. Copyright © 2010 John Wiley & Sons, Ltd.

Received 28 March 2007; Revised 3 December 2009; Accepted 12 February 2010

KEY WORDS: health-related quality of life; preference-based index measurement; methodology; utility of health states; statistical approach

1. INTRODUCTION
Health-economic evaluations are often concerned with health-related quality of life (HRQoL). Especially, HRQoL in the sense of health-state valuations given by the average health care system user is of interest. Usually, to measure HRQoL in this sense, preference-based index measurement instruments are applied. Each instrument of this kind consists of (1) a multi-attribute classification system for distinguishing health states and (2) a scoring function which assigns a valuation to each health state distinguished by the classification system. Usually, the classification system is defined via questions presented in a questionnaire, whereas the scoring function is given as a part of the instructions for evaluating the completed questionnaires. The best known examples of preference-based index measurement instruments are the Quality of Well-Being Scale (Kaplan et al., 1976), the Health-Utilities-Index (Feeny et al., 1995), the EQ–5D (Brooks 1996), the SF–6D (Brazier et al., 2002,1998), and the 15D (Sintonen, 2001).

*Correspondence to: Trimberg Research Academy, University of Bamberg, St.-Getreu-Straße 14, 96049 Bamberg, Germany. E-mail: uwe.konerding@uni-bamberg.de

Copyright © 2010 John Wiley & Sons, Ltd.
To construct a preference-based index measurement instrument, all health states distinguished by the classification system must be valuated. Yet, even classification systems with only a few attributes can distinguish a considerable number of health states. Therefore, very often, scoring functions are determined by valuating only a small sub-set of the distinguished health states directly and by computing the scoring function for all health states from the empirically obtained values using a theoretical model (Froberg and Kane, 1989). This approach is sometimes referred to as the statistical approach. The theoretical model applied within this approach will here be referred to as an index model. Hitherto, the statistical approach has been applied for determining scoring functions for the EQ–5D (Busschbach et al., 1999; Claes et al., 2002; Dolan, 1997; Johnson et al., 1998; Kerim-Sade and Wasem, 2000; Schulenburg et al., 1996, 1998; Tsuchiya et al., 2002) and for the SF–6D (Brazier et al., 2002, 1998).

The statistical approach is highly ingenious. It is, however, also associated with problems. These problems arise because the empirically obtained health-state valuations contain arbitrary settings. These settings affect the final scoring function values. The manner in which they do this depends on the index model applied for computation. With an inadequate model, the scoring function values mainly reflect arbitrary settings. In this case, they hardly form an adequate basis for serious health economical analyses. For this purpose, scoring functions are needed which predominantly reflect the preferences of the originally questioned persons. Therefore, in this article, the role of arbitrary settings in determining scoring functions is investigated more closely. The complete argumentation is divided into five parts: (a) the three most important techniques for directly valuating health states are outlined and the manner in which they are based on arbitrary settings is elaborated; (b) some general conceptions for handling arbitrary settings in measurement values are introduced; (c) the most important models applied as index models are presented and the way in which they interact with arbitrary settings is analysed; (d) the results of these analyses are demonstrated by computational examples; and (e) implications concerning further research are discussed.

2. DIRECT VALUATION TECHNIQUES AND ARBITRARY SETTINGS

Predominantly, three procedures are applied as direct valuation techniques for HRQoL: the standard gamble (SG), the time-trade-off (TTO) and the rating-scale procedure (RSP). SG values are mostly referred to as utilities, whereas TTO and RSP values are rather referred to as valuations. In this context, the term ‘valuation’ is applied to all three techniques, which will now be discussed.

2.1. Standard gamble

In SG, subjects are asked to choose between two alternative prospects. For target health states better than death, the first prospect is living in this health state; the second is undergoing medical treatment with the two alternative uncertain outcomes, death and full health. The probabilities for the two alternative uncertain outcomes are changed until the subject judges both prospects as equivalent. The probability assigned to full health is then taken as a measurement value of the target health state. For target health states worse than death, the first prospect is usually certain death. The second prospect is undergoing medical treatment with two uncertain outcomes: (1) full health and (2) the target health state. Again, the probabilities of the uncertain outcomes are changed until the subject judges both prospects as equivalent. The measurement value for the target health state is then determined as \( -\frac{p}{1-p} \) with \( p \), the probability of health in the second prospect.

SG measurements are based on three components: (1) empirically obtained judgments, (2) a theoretical assumption, and (3) arbitrary settings. The empirically obtained judgments are the subjects’ preference and equivalence judgments concerning the two alternative prospects. The theoretical assumption is the model of subjectively expected utility (SEU-model) originally proposed by Bernoulli (1738) and later on axiomatised by Von Neumann and Morgenstern (1944). According to this model,
the valuation of the whole prospect is a linear combination of the valuations and the probabilities of the single outcomes, i.e.

\[ v(X) = \sum_{i=1}^{n} v(x_i) \times p(x_i) \]  

(1)

where \( v(X) \) is the valuation of the whole prospect, \( n \) the number of the different possible outcomes, \( v(x_i) \) the valuation of outcome \( i \), and \( p(x_i) \) the corresponding probability. The arbitrary settings consist in setting the valuations of full health and death equal to one and zero, respectively. Only because of these settings can the target health state’s valuation be determined in the manner described above.

The two arbitrary settings incorporated in the SG are so generally accepted that they might seem to be not arbitrary at all. However, they are. The empirical information as given by the subjects and as interpreted by the SEU-model could just as well be represented numerically if the valuations of full health and death were coded with 100 and with −100, respectively. In this case, only the rule for coding the target health state would need to be adapted, i.e. the coding rule would have to be \( 200^\ast(x-0.5) \) with \( x \), the value obtained from the standard coding with 1 for full health and 0 for death. In principle, any coding with a higher value for full health than for death is possible. However, all these codings are much less convenient than the standard coding and hence there is no good reason to apply them.

2.2. Time-trade-off

In TTO, subjects are also confronted with two prospects concerning health. For target health states better than death, the first prospect is living a certain time in the target health state and then dying immediately. The second prospect is living a shorter time in full health and then dying immediately. The time of living in full health is changed until the subject judges both prospects as equivalent. The valuation of the target health state is then determined by \( d_{th}/d_{ths} \) with \( d_{th} \) the duration of full health within the second prospect and \( d_{ths} \) the duration of the target health state within the first prospect. For target health states worse than death, the first prospect is dying immediately; the second consists of first living a certain time in the target health state and then living a certain time in full health. Both durations in the second prospect are changed until the subject judges both prospects as equivalent. The valuation of the target health state is then determined as \(-d_{th}/d_{ths}\) with \( d_{ths} \) the duration of the target health state and \( d_{th} \) the duration of full health, both within the second prospect.

Just like SG measurements, TTO measurements are based on three components: (1) empirically obtained judgments, (2) a theoretical assumption, and (3) arbitrary settings. Again, the empirically obtained judgments are the preference and equivalence judgments concerning the two prospects. The theoretical assumption is less obvious than in the case of the SG because this assumption is hardly ever explicitly stated. However, it is always implicitly presupposed. According to this assumption the valuation of the whole sequence of states is a linear combination of the valuations and durations of the single states constituting this sequence, i.e.

\[ v(X) = \sum_{i=1}^{n} v(x_i) \times d(x_i) \]  

(2)

where \( v(X) \) is the valuation of the whole sequence, \( n \) the number of the different time intervals constituting this sequence, \( v(x_i) \) the valuation of the health state belonging to interval \( i \), and \( d(x_i) \) the corresponding duration. The arbitrary settings consist in setting the valuation of health equal to 1 and the valuation of death equal to 0. Only because of these settings, the valuation of the target health state can be determined according to the rules described above.

As in SG, in TTO, the empirical information as given by the subjects and as interpreted by the model could also just as well be numerically represented using any two other numbers for full health and death with the number for full health higher than the number for death. In this case, only the formula for
computing the valuation of the target health state would have to be adapted. For example, if full health
were represented by 100 and death by \(-100\), then the value for the target health state would be
\(200^*(x-0.5)\) with \(x\) the value obtained from the standard coding with 1 for full health and 0 for death. Admittedly, this would be very awkward and no attempt is made to propose this scaling. Nevertheless, the measurement values produced by this scaling would represent exactly the same empirical meaning as the measurement values with standard coding.

2.3. Rating-scale procedure

In RSP, subjects are either confronted with a straight line, i.e. a visual analogue scale, or a set of
categories with equal intervals between the categories, i.e. a categorical scale. Both ends of the scales are
linked with words denoting health states. Usually, one of these health states is full health; the other is,
in most cases, the worst conceivable health state. Subjects are asked to mark a point on the scale so
that the distances between this mark and the two extremes of the scale correspond to the distances
between the target health state’s valuation and the valuations of the two reference states. To obtain a
numerical value for the target health state’s valuation, numbers are assigned to both extreme health
states, usually with a higher number for full health, and the other positions on the scale are numerically
represented so that the distances between the numbers correspond to the distances between the positions
on the scale. The scale value, which belongs to the position where the subject has placed his or her mark,
is then taken as the valuation of the target health state. When death has also been rated, the resulting
scale is often linearly transformed with 1 assigned to full health and 0 to death.

Just like in SG and TTO, RSP measurements are based on three components: (1) empirically
obtained judgments, (2) a theoretical assumption, and (3) arbitrary settings. The empirically obtained
judgements are the marks on the scale. The theoretical assumption is that the subjects place the mark on
the scale so that the intervals between the mark and the two scale endpoints correspond to the intervals
between the corresponding valuations, i.e.

\[
\frac{s(x) - s(x_1)}{s(x_u) - s(x_1)} = \frac{v(x) - v(x_1)}{v(x_u) - v(x_1)}
\]

where \(x\) is the target health state, \(x_1\) the health state at the lower, and \(x_u\) the health state at the upper
scale endpoint, \(s\) the values on the rating scale and \(v\) the actual valuations. The arbitrary settings are the
numbers assigned to the two scale endpoints or, in the case of a linear transformation, the two numbers
assigned to full health and death. In contrast to the other two techniques, all these settings are quite
obviously arbitrary.

3. CONCEPTIONS FOR HANDLING ARBITRARY SETTINGS

All sets of measurement values that can be produced by changing the arbitrary settings in SG, TTTO, or
RSP are systematically related to each other. They can be transformed into each other by functions of
the type \(f(x) = \beta^*x + \alpha\) with a positive \(\beta\). In measurement theory (Krantz et al., 1971; Luce et al., 1990;
Suppes et al., 1989; Suppes and Zinnes, 1963), transformations by means of which different yet
empirically equivalent sets of measurement values can be transformed into each other are referred to as
permissible transformations. The less a measurement depends upon arbitrary settings the smaller is the
class of permissible transformations. A set of measurement values with permissible transformations of
the type just described is referred to as an interval scale. Hence, under presupposition of the theoretical
assumptions described above, SG, TTTO, and RSP produce interval scales.

The concept of permissible transformations makes it possible to analyse exactly how the arbitrary
settings in the basic measurements, i.e. in the direct measurements of HRQoL, affect the scoring
function when this function is computed according to the statistical approach. The central idea is to
investigate how the scoring function changes when a permissible transformation is applied to the basic measurements. In accordance with the two computational steps of the statistical approach, this analysis is divided into two parts. The first is concerned with the effects of the permissible transformations on the model parameters and the second with the effect of the resulting changes in the model parameters on the final scoring function.

The effect of the permissible transformations on the model parameters can be checked by five steps:

1. The model equation must be formulated for the original values \( v(y_p) \) with the corresponding model parameters \( \varphi_1, \varphi_2, \ldots, \varphi_\Omega \).
2. The model equation must be formulated for the transformed measurement values \( v'(y_p) \) with the corresponding model parameters \( \varphi'_1, \varphi'_2, \ldots, \varphi'_\Omega \).
3. The term \( v'(y_p) \) in the equation of the transformed values must be substituted according to \( v'(y_p) = \beta v(y_p) + \alpha \).
4. The resulting equation must be solved for \( v(y_p) \).
5. The equation resulting from the fourth step must be compared with the equation for the original values.

The fifth step directly renders the relationships between the parameters estimated for the two different sets of measurement values. Once the relationships between the two sets of model parameters have been elaborated, the interrelationships between the two corresponding scoring functions can be determined directly.

When the basic measurement values possess interval scale level, the same should hold for the resulting scoring function. In other words, the scoring function belonging to the original measurement values should be related to the scoring function produced from permissibly transformed basic measurement values by

\[
\beta * S_{\text{MOD}}[v(y_p)] + \alpha = S_{\text{MOD}}[v'(y_p)]
\]

with \( S_{\text{MOD}}[v(y_p)] \) the scoring function determined using the respective index model on the basis of the untransformed measurement values, \( S_{\text{MOD}}[v'(y_p)] \) the corresponding function for the transformed values, and \( \beta \) positive. This condition should be fulfilled by each index model. Otherwise, the ratios between the intervals of the resulting scoring function vary with the arbitrary settings applied for coding death and full health. Consequently, in this case these intervals would have no reasonable empirical meaning and the corresponding scoring functions would not be appropriate for determining incremental cost-effectiveness ratios and would therefore not be appropriate for evaluating medical interventions.

4. FORMAL EVALUATION OF SOME PARADIGMATIC MODELS

Several models have been proposed as index models. In most of these models, the valuations of the whole health states are predicted by valuations of specific aspects of these states. Two classes of models can be distinguished: (1) additive and (2) multiplicative models. Each additive model has a multiplicative counterpart and vice versa. In the following, first the most typical additive and then the analogous multiplicative models are analysed with respect to their interaction with arbitrary settings in the basic measurements.

4.1. Additive models

The most fundamental additive model is

\[
v(y_p) = c + \sum_{i=1}^{n} \sum_{r=2}^{m(i)} w(x_{ir}) * d(x_{ir}, y_p)
\]
with \( v(y_p) \) the valuation of health state \( y_p \), \( c \) an additive parameter, \( n \) the number of attributes constituting the multi-attribute classification system, \( m(i) \) the number of attribute levels of attribute \( i \), \( w(x_{ir}) \) the valuation of level \( r \) of attribute \( i \), and \( d(x_{ir}, y_p) \) a dummy variable which is one if level \( r \) of attribute \( i \) belongs to \( y_p \) and otherwise zero. The terms \( \gamma \) and \( w(x_{ir}) \) must be estimated from the basic measurements. The valuation of the first level of each attribute is implicitly set equal to 0. In the following, the model will be referred to as the paradigmatic additive model with free additive parameter (PAM-free). Hitherto, this model has been applied by Brazier et al. (1993) and by Busschbach et al. (1999, Equation (7)).

The equation for the transformed basic measurements is

\[
v'(y_p) = c' + \sum_{i=1}^{n} \sum_{r=2}^{m(i)} w'(x_{ir}) \times d(x_{ir}, y_p)
\]  

\(6\)

Substituting according to \( v'(y_p) = \beta \times v(y_p) + \alpha \) and solving for \( v(y_p) \) renders

\[
v(y_p) = (c' - \alpha)/\beta + \sum_{i=1}^{n} \sum_{r=2}^{m(i)} [w'(x_{ir})/\beta] \times d(x_{ir}, y_p)
\]  

\(7\)

Comparing the right-hand terms in Equations 5 and 7 reveals that equations

\[
c = (c' - \alpha)/\beta
\]  

\(8\)

and

\[
w(x_{ir}) = w'(x_{ir})/\beta
\]  

\(9\)

hold.

The scoring function for the original measurement values, i.e. \( S_{MOD}[v(y_p)] \), is computed using

\[
S_{MOD}[v(y_p)] = c + \sum_{i=1}^{n} \sum_{r=2}^{m(i)} w(x_{ir}) \times d(x_{ir}, y_p)
\]  

\(10\)

and the scoring function for the transformed measurement values, i.e. \( S_{MOD}[v'(y_p)] \), using

\[
S_{MOD}[v'(y_p)] = c' + \sum_{i=1}^{n} \sum_{r=2}^{m(i)} w'(x_{ir}) \times d(x_{ir}, y_p)
\]  

\(11\)

Substituting in Equation 10 according to Equations 8 and 9 and rearranging renders

\[
\beta \times S_{MOD}[v(y_p)] + \alpha = c' + \sum_{i=1}^{n} \sum_{r=2}^{m(i)} w'(x_{ir}) \times d(x_{ir}, y_p)
\]  

\(12\)

Equations 11 and 12 imply

\[
\beta \times S_{MOD}[v(y_p)] + \alpha = S_{MOD}[v'(y_p)]
\]  

\(13\)

i.e. the condition described in Equation 4.

PAM-free can be modified by adding interaction terms. For example, a term which expresses whether at least one of the attributes is at its worst level can be added. Such a model has been applied by Dolan (1997) and Johnson et al. (1998). PAM-free can also be modified by imposing restrictions upon the model parameters. For example, the intervals between the different levels of the same attribute can be assumed to be equal. Such a model has been applied by Busschbach et al. (1999, Equation (9)). Performing the same investigation as presented above shows that the condition described in Equation 4 holds for all modifications just mentioned.
PAM-free can further be modified by setting the free additive parameter equal to a constant value. If this value is 0, the corresponding equation is

\[
v(y_p) = \sum_{i=1}^{n} \sum_{r=2}^{m(i)} w(x_{ir}) * d(x_{ir}, y_p)
\] (14)

In the following, this model will be referred to as the paradigmatic additive model with the additive parameter set equal to zero (PAM-zero). Rewriting the model equation for the transformed measurement values and the corresponding estimated model parameters renders

\[
v'(y_p) = \sum_{i=1}^{n} \sum_{r=2}^{m(i)} w'(x_{ir}) * d(x_{ir}, y_p)
\] (15)

Substituting according to \(v'(y_p) = \beta^* v(y_p) + \alpha\) and solving for \(v(y_p)\) renders

\[
v(y_p) = -\alpha/\beta + \sum_{i=1}^{n} \sum_{r=2}^{m(i)} w'(x_{ir})/\beta * d(x_{ir}, y_p)
\] (16)

Comparing the right-hand sides of Equations 14 and 16 reveals a substantial structural difference. In contrast to Equation 14, Equation 16 contains the additive term \(-\alpha/\beta\). This term is different from zero if \(\alpha\) is different from zero.

As Equation 16 is equivalent to Equation 15, the structural difference between Equations 14 and 16 implies that the model usually fits differently well to the measurement values which differ only with respect to an additive transformation. Moreover, this structural difference implies that there is no uniquely defined relationship between the parameters estimated for \(v(y_p)\) and those estimated for \(v'(y_p)\). This relationship depends not only on the parameters of the permissible transformation but also on data. Consequently, there is no systematic relationship between the two corresponding scoring functions. Therefore, PAM-zero produces no empirically meaningful scoring function with the basic measurement values applied here. This result will be the same for all additive models for which the additive parameter is set equal to an \textit{a priori} given number.

### 4.2. Multiplicative models

The multiplicative counterpart for PAM-free is

\[
v(y_p) = e * \prod_{i=1}^{n} \prod_{r=2}^{m(i)} w(x_{ir}) * d(x_{ir}, y_p)
\] (17)

In the following, this model will be referred to as the paradigmatic multiplicative model with free multiplicative parameter (PMM-free). The terms \(v(y_p)\), \(n\), \(m(i)\), \(w(x_{ir})\) and \(d(x_{ir}, y_p)\) are defined as above. Also, just like in PAM-free, \(e\) is a parameter which must be estimated from the basic measurements. However, unlike in PAM-free, this parameter is multiplicative. Just like in the additive models, this parameter must be estimated from data together with the \(w(x_{ir})\), and, just like in PAM-free, the valuation for the first level of each attribute is set implicitly. However, in contrast to PAM-free, it is set equal to 1 and not equal to 0. Up to now, PMM-free has been applied by Kerim-Sade and Wasem (2000) and, with slight variations, by Schulenburg \textit{et al.} (1996, 1998).

Applying the model to the transformed measurement values renders

\[
v'(y_p) = e^* \prod_{i=1}^{n} \prod_{r=2}^{m(i)} w'(x_{ir}) * d(x_{ir}, y_p)
\] (18)
Substituting according to \( v'(y_p) = \beta v(y_p) + \alpha \) and solving for \( v(y_p) \) renders

\[
v(y_p) = \frac{-\alpha}{\beta} + c^* \prod_{i=1}^{n} \prod_{r=2}^{m(r)} \left[w(x_{ir})^{d(x_{ir}, y_p)}\right] / \beta
\]  

(19)

Comparing the right-hand sides of Equations 17 and 19 reveals an essential structural difference. In contrast to Equation 17, Equation 19 contains the additive term \(-\alpha/\beta\). This term is different from zero if \( \alpha \) is different from zero.

The structural difference between Equations 17 and 19 implies that PMM-free fits differently well to sets of measurement values which differ only with respect to an additive transformation. Moreover, the structural difference implies that there is no uniquely defined relationship between the parameters estimated on the basis of two data sets related in this way. Hence, neither the corresponding scoring functions are related systematically. This in turn means that PMM-free does not produce empirically meaningful scoring functions for the kind of basic measurement values considered here. Investigating in an analogous way the multiplicative counterparts of the other additive models discussed above produces the same result.

5. COMPUTATIONAL DEMONSTRATION

The effects of the model on the resulting scoring function will now be demonstrated with the data applied by Dolan (1997) to determine the scoring function for the EQ-5D. These data stem from 3395 persons and contain 39,953 valuations referring to 42 different health states. Four different models are considered: (1) PAM-free, (2) PAM-zero, (3) PMM-free, and (4) the paradigmatic multiplicative model with the multiplicative parameter set equal to one (PMM-one) (Table I). To investigate the effects of changing the arbitrary settings in the basic measurements, these models were applied to nine different scales. One is the original scale with the valuations of death and full health set equal to 0 and 1, respectively. The other eight scales were produced by systematically combining additive and multiplicative transformations performed with the original scale (Table II). Correspondingly, the other eight scales are no longer standardised with the valuations of death and full health set equal to 0 and 1.

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAM-free</td>
<td>( v(y_p) = c + \sum_{i=1}^{n} \sum_{r=2}^{m(r)} w(x_{ir}) \cdot d(x_{ir}, y_p) )</td>
</tr>
<tr>
<td>PAM-zero</td>
<td>( v(y_p) = \sum_{i=1}^{n} \sum_{r=2}^{m(r)} w(x_{ir}) \cdot d(x_{ir}, y_p) )</td>
</tr>
<tr>
<td>PMM-free</td>
<td>( v(y_p) = c \prod_{i=1}^{n} \prod_{r=2}^{m(r)} w(x_{ir}) \cdot d(x_{ir}, y_p) )</td>
</tr>
<tr>
<td>PMM-one</td>
<td>( v(y_p) = \prod_{i=1}^{n} \prod_{r=2}^{m(r)} w(x_{ir}) \cdot d(x_{ir}, y_p) )</td>
</tr>
</tbody>
</table>

Table II. The nine test scales

<table>
<thead>
<tr>
<th>Additive transformation</th>
<th>Multiplicative transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( -1 )</td>
<td>( A = 0.5^* (OS - 1) )</td>
</tr>
<tr>
<td>( \pm 0 )</td>
<td>( D = 0.5^* OS )</td>
</tr>
<tr>
<td>( +1 )</td>
<td>( G = 0.5^* (OS + 1) )</td>
</tr>
</tbody>
</table>

\( ^* \) OS, original scale.
For all four models and for all nine scales, the parameters were estimated by means of least-square-estimation with respect to the original model equation, i.e. the multiplicative models were not logarithmised before parameter estimation. To judge the fit-to-data, the adjusted multiple $R^2$, i.e. $1 - (\text{mean squared residuals})/(\text{variance of predicted values})$, was computed for all four models and for all nine scales. Subsequently, for each model and each scale, a scoring function was computed for all 243 EQ-5D-health-states. To investigate how the transformations affect the scoring functions, the Pearson correlations between the different scoring functions belonging to the same model were computed. If the condition described in Equation 4 holds, these correlations must be one between all scoring functions belonging to the same model. To check how changes in the arbitrary settings affect the rank order of the scoring function values, Kendall’s tau was also computed for each pair of scales determined with the same model.

The results correspond to the mathematical analyses. The $R^2$ statistics for the PAM-free do not change in dependence on the scale transformations. For all of the other three models, however, they do change (Table III). For PAM-zero and PMM-free, they change with additive transformations of the basic measurements; for PMM-one they change both with additive and multiplicative transformations of the basic measurements. The $R^2$ statistics also reflect that PMM-one is not suited for adjusting changes in the magnitude of the criterion values. For this model, $R^2$ even becomes negative in some cases. The results for the Pearson correlations also correspond to the mathematical analyses. For PAM-free, these correlations are all equal to 1. For the other three models, this is not true (Table IV, upper right halves). For PAM-zero and PMM-one, additive transformations of the basic measurements produce correlations lower than 1. For PMM-one none of the correlations is 1. There are even negative correlations. For the rank correlations, the pattern is by and large the same (Table IV, lower left halves).

### 6. IMPLICATIONS FOR FURTHER RESEARCH

The considerations just presented have one central implication for further research: if scoring functions are determined from SG, TTO, or RSP values using the statistical approach, then only additive models with a freely estimated additive parameter should be applied. Otherwise, scoring functions with arbitrarily mixed up intervals and even arbitrarily mixed up rank orders are produced. Correspondingly, scoring functions which have already been constructed by means of different models should not be applied because the intervals and the rank orders of these scoring functions are essentially affected by arbitrary settings. Finally, working with scoring functions produced with inadequate index models will lead to inappropriate estimations of incremental cost effectiveness ratios and thereby to inappropriate judgements concerning the cost effectiveness of new medical interventions.

The class of additive models with a free additive parameter is very large. In principle, each empirical structure in the health-state valuations can be modelled with a model from this class. Models which are

<table>
<thead>
<tr>
<th>Scale</th>
<th>Paradigmatic additive models</th>
<th>Paradigmatic multiplicative models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PAM-free</td>
<td>PAM-zero</td>
</tr>
<tr>
<td>A</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>B</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>C</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>D</td>
<td>0.32</td>
<td>0.31</td>
</tr>
<tr>
<td>E</td>
<td>0.32</td>
<td>0.31</td>
</tr>
<tr>
<td>F</td>
<td>0.32</td>
<td>0.31</td>
</tr>
<tr>
<td>G</td>
<td>0.32</td>
<td>0.28</td>
</tr>
<tr>
<td>H</td>
<td>0.32</td>
<td>0.28</td>
</tr>
<tr>
<td>I</td>
<td>0.32</td>
<td>0.28</td>
</tr>
</tbody>
</table>
Table IV. Correlations between scoring functions

<table>
<thead>
<tr>
<th>Scale</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>I</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAM-free</td>
<td>—</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>B</td>
<td>1.00</td>
<td>—</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>C</td>
<td>1.00</td>
<td>1.00</td>
<td>—</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>D</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>—</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>E</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>—</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>F</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>—</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>G</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>—</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>H</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>—</td>
<td>1.00</td>
</tr>
<tr>
<td>I</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>—</td>
</tr>
<tr>
<td>PAM-zero</td>
<td>—</td>
<td>1.00</td>
<td>1.00</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>B</td>
<td>1.00</td>
<td>—</td>
<td>1.00</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>C</td>
<td>1.00</td>
<td>1.00</td>
<td>—</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>D</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>—</td>
<td>1.00</td>
<td>1.00</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>E</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>1.00</td>
<td>—</td>
<td>1.00</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>F</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>1.00</td>
<td>1.00</td>
<td>—</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>G</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>—</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>H</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>1.00</td>
<td>—</td>
<td>1.00</td>
</tr>
<tr>
<td>I</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>1.00</td>
<td>1.00</td>
<td>—</td>
</tr>
<tr>
<td>PMM-free</td>
<td>—</td>
<td>1.00</td>
<td>1.00</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.51</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>B</td>
<td>1.00</td>
<td>—</td>
<td>1.00</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.51</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>C</td>
<td>1.00</td>
<td>1.00</td>
<td>—</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.51</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>D</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>—</td>
<td>1.00</td>
<td>1.00</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>E</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>1.00</td>
<td>—</td>
<td>1.00</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>F</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>1.00</td>
<td>1.00</td>
<td>—</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>G</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>—</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>H</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>1.00</td>
<td>—</td>
<td>1.00</td>
</tr>
<tr>
<td>I</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.34</td>
<td>0.34</td>
<td>0.34</td>
<td>1.00</td>
<td>1.00</td>
<td>—</td>
</tr>
<tr>
<td>PMM-one</td>
<td>—</td>
<td>0.70</td>
<td>0.33</td>
<td>0.01</td>
<td>–0.03</td>
<td>0.66</td>
<td>0.01</td>
<td>0.00</td>
<td>–0.00</td>
</tr>
<tr>
<td>B</td>
<td>0.81</td>
<td>—</td>
<td>0.65</td>
<td>0.16</td>
<td>-0.16</td>
<td>0.99</td>
<td>0.18</td>
<td>0.08</td>
<td>0.02</td>
</tr>
<tr>
<td>C</td>
<td>0.61</td>
<td>0.73</td>
<td>—</td>
<td>0.20</td>
<td>-0.28</td>
<td>0.68</td>
<td>0.18</td>
<td>0.15</td>
<td>0.12</td>
</tr>
<tr>
<td>D</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.00</td>
<td>—</td>
<td>-0.02</td>
<td>0.17</td>
<td>0.96</td>
<td>0.94</td>
<td>0.78</td>
</tr>
<tr>
<td>E</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.04</td>
<td>-0.03</td>
<td>—</td>
<td>-0.19</td>
<td>0.02</td>
<td>0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>F</td>
<td>0.78</td>
<td>0.97</td>
<td>—</td>
<td>-0.01</td>
<td>0.02</td>
<td>—</td>
<td>0.19</td>
<td>0.08</td>
<td>0.02</td>
</tr>
<tr>
<td>G</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.00</td>
<td>0.91</td>
<td>-0.03</td>
<td>-0.01</td>
<td>—</td>
<td>0.88</td>
<td>0.68</td>
</tr>
<tr>
<td>H</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.00</td>
<td>0.85</td>
<td>-0.03</td>
<td>-0.01</td>
<td>0.82</td>
<td>—</td>
<td>0.93</td>
</tr>
<tr>
<td>I</td>
<td>-0.00</td>
<td>-0.00</td>
<td>0.00</td>
<td>0.71</td>
<td>-0.02</td>
<td>-0.00</td>
<td>0.67</td>
<td>0.84</td>
<td>—</td>
</tr>
</tbody>
</table>

aRight upper halves of the matrices: Pearson’s rho; left lower halves of the matrices: Kendall’s tau-b.

Empirically less restrictive than PAM-free can be produced by adding various interaction terms. Nevertheless, in spite of the broad possibilities given by the additive models with a free additive parameter, the unrelenting demand to use only models of this kind contradicts, in fact or only apparently, some practices of this research. In the following, two of the most important of these practices are discussed. One is inside and the other is outside the statistical approach.

6.1. Research inside the statistical approach

For the computation of quality adjusted life years (QALYs), scoring functions are required which are standardised with the value for death equal to 0 and the value for full health equal to 1. In the statistical approach, such a standardisation cannot be produced by simply starting with basic measurement values which are already standardised in the desired way. Because of certain statistical regularities of multiple
linear regression analyses (Greene, 1997, Chapter 6), scoring functions values which have been
produced with an additive model with a free additive parameter will be lower for the better and higher
for the worse health states in comparison with the corresponding basic measurement values. This effect
has already been reported by Brazier et al. (2002) and Dolan (1997).

To obtain the desired standardisation in spite of the complications just described, Brazier et al. (2002)
start with basic measurements in the desired standardisation and apply an additive model for which the
additive parameter is set a priori so that the scoring function value for full health is equal to 1. Actually,
this approach guarantees that the scoring function values for the better health states are no longer lower
than the original standardised measurement values for these health states. However, in this approach,
the scoring function values for the worse health states are still higher than the corresponding basic
measurement values. Moreover, according to the considerations presented above, this approach results
in an arbitrary mixing up of the intervals and even the rank order of the final scoring function values.
The problems just described can be solved by applying the following two-step approach:

1. The scoring function is estimated by means of an additive model with additive parameter and on
the basis of basic measurement values which are standardised so that 1 represents full health and
0 death.
2. The resulting scoring is standardised again with 1 for full health and 0 for death.

As soon as both death and full health can be specified within the classification system, the final
standardisation of the scoring function can be performed simply by applying a transformation of the
type \( f(x) = \beta x + z \) which maps the unstandardised scoring function value of death on 0 and the
unstandardised scoring function value of full health on 1. Unfortunately, in most classification systems
only full health can be specified, but not death. Hence, a different approach must be applied. With the
basic measurements standardised in the desired way, the final standardisation can best be done by
mapping the unstandardised scoring function value for full health on 1 and by forcing the standard
deviation of the standardised scoring function values to be equal to the variance of the basic
measurements.

Let \( \text{SD} \left( \text{BM}_t \right) \) be the standard deviation of the standardised basic measurements, \( \text{SD}(x) \) the standard
deviation of the unstandardised scoring function values, \( x(t) \) the unstandardised scoring function value
of health state \( t \) and \( x(fh) \) the unstandardised scoring function value for full health, then the formula for
producing the standardised scoring value for \( t \), i.e. \( y(t) \), is

\[
y(t) = 1 - \frac{\text{SD}(\text{BM}_t) \times [x(fh) - x(t)]}{\text{SD}(x)}
\]  

(20)

This formula produces scoring function values with the value for full health equal to one and with
values for the worse health states in the same magnitude as the corresponding basic measurement
values. The latter implies that this kind of standardisation better guarantees that zero is in fact the
valuation for death than is the case in the approach applied by Brazier et al. (2002).

6.2. Research outside the statistical approach

The statistical approach is not the only approach concerned with determining valuations for multi-
attribute objects. This is also the concern of prescriptive decision theory for multi-attribute objectives
(Keeney and Raiffa, 2003). In the context of this theory, three functions for predicting the utility of a
multi-attribute objective using single attribute utilities have been presented: the additive, the
multiplicative, and the multi-linear utility function (Keeney and Raiffa, 2003, Chapter 6). Usually,
these functions are referred to as multi-attribute utility (MAU) functions. None of these functions
contains an additive parameter and, in the multiplicative as well as in the multi-linear function, single
attribute utilities are multiplied with each other.

The MAU-functions might seem to contradict the demand formulated above. In fact, they do not. This
is so because the scaling procedure to which the MAU-functions refer differs essentially from the procedure
applied in the statistical approach. In the MAU-scaling the multi-attribute utilities, the single attribute
utilities and the model parameters are scaled together at the same time and in such a manner that several
specific relationships between the different utilities and the model parameters hold. These relationships
imply that no additive parameters are required and that multiplying utilities with utilities makes sense.
These relations, however, cannot be established for certain in the statistical approach. Therefore, within
the statistical approach, the MAU-functions cannot be applied without extensive modifications. To reconcile
both approaches, those additive models with additive parameters should be specified which correspond to
the three MAU-functions. This, however, must be subjected to further research.

ACKNOWLEDGEMENTS

This work was funded by the Ministry for Education and Research of the Federal Republic of Germany
(grants no. 01ZZ0103 and 01ZZ0403). The author thanks Gabriele Klever-Deichert, Thomas
Kohlmann, Jörn Moock, Ute Schmid and two anonymous reviewers for critically discussing former
versions of the manuscript, Paul Kind for providing the data for the EQ-5D-study, and Peter Bereza for
correcting the author’s English.

REFERENCES

Bernoulli D. 1738. Specimen theoriae novae de mensura sortis. Comentarii Academiae Scientiarum Imperiales
Brazier J, Jones N, Kind P. 1993. Testing the validity of the EuroQol and comparing it with the SF 36 health survey
health description and health valuation with an application to the EuroQol EQ–5D. Journal of Health
Economics 18: 551–571.
Claes C, Greiner W, Über A. 2002. Der EQ-5D (EuroQol) als krankheitsübergreifendes Indexinstrument (The EQ-
5D (EuroQol) as a generic index instrument). In Gesundheitsökonomische Evaluationen (Health-Economical
University Press: Cambridge, UK.


