Pergamon

PII: S0277-9536(97)00099-3

Soc. Sci. Med. Vol. 45, No. 11, pp. 1641–1652, 1997

Printed in Great Britain
0277-9536/97 $17.00 + 0.00

THE COMPARABILITY AND RELIABILITY OF FIVE HEALTH-STATE VALUATION METHODS

PAUL F. M. KRABBE*, MARIE-LOUISE ESSINK-BOT and GOUKE J. BONSEL

Department of Public Health, Faculty of Medicine, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Abstract—The objective of the study was to consider five methods for valuing health states with respect to their comparability (convergent validity, value functions) and reliability. Valuation tasks were performed by 104 student volunteers using five frequently used valuation methods: standard gamble (SG), time trade-off (TTO), rating scale (RS), willingness-to-pay (WTP) and the paired comparisons method (PC). Throughout the study, the EuroQol classification system was used to construct 13 health-state descriptions. Validity was investigated using the multitrait–multimethod (MTMM) methodology. The extent to which results of one method could be predicted by another was examined by transformations. Reliability of the methods was studied parametrically with Generalisability Theory (an ANOVA extension), as well as non-parametrically. Mean values for SG were slightly higher than TTO values. The RS could be distinguished from the other methods. After a simple power transformation, the RS values were found to be close to SG and TTO. Mean values of WTP were linearly related to SG and TTO, except at the extremes of the scale. However, the reliability of WTP was low and the number of inconsistencies substantial. Valuations made by the RS proved to be the most reliable. Paired comparisons did not provide stable results. In conclusion, the results of the parametric transformation function between RS and SG/TTO provide evidence to justify the current use of RS (with transformations) not only for reasons of feasibility and reliability but also for reasons of comparability. A definite judgement on PC requires data of a complete design. Due to the specific structure of the correlation matrix which is inherent in valuing health states, we believe that full MTMM is not applicable for the standard analysis of health-state valuations. © 1997 Elsevier Science Ltd. All rights reserved

Key words—valuation methods, validity, comparability, reliability, methodology, EuroQol

INTRODUCTION

It is no longer sufficient to gather data on mortality and medical disease-specific parameters to evaluate the burden of disease and the effects of medical interventions from a societal perspective. Data on economic costs and on health status is also required. Considerable effort has been invested in the development of general indicators which allow for valid comparison of health status effects across different diseases. It is generally agreed that such indicators should be derived from a comprehensive concept of health, covering at least the physical, psychological and social domains. Several indicators are currently available, some of them suitable for use in economic cost-utility analysis.

The following three-stage procedure is frequently used to incorporate health status effects in utility analyses (Brooks, 1995; Essink-Bot, 1995). In stage I, the course of a disease is divided into broadly homogenous phases and patients' health status in each phase is measured using a descriptive system. In stage II, the health status descriptions that correspond to the disease phases are formally valued. Results from stage I and II can then be combined with duration data in stage III to calculate quality-adjusted life-years as an outcome measure.

The valuation of health states (stage II), forms a critical part of this three-stage approach. Several valuation methods (methodologically labelled: scaling methods; Froberg and Kane, 1989a) exist, each with their own theoretical framework and conceptual position. We investigated five established health-state valuation methods. First, we looked at a common rating scale, a seemingly simple method. Second, we investigated two economic methods, standard gamble (considered to be the approximate operationalisation of game theory) and willingness-to-pay, each referred to as trade-off methods. From an economic point of view willingness-to-pay can be considered to be the superior quantification of non-monetary aspects of disease (Thompson *et al.*, 1982; Thompson *et al.*, 1984; Gafni, 1991; O'Brien and Gafni, 1996). We also investigated another trade-off method, the time trade-off. This method occupies a position in between, i.e. being considered as more feasible than standard gamble and more

*Author for correspondence. Department of Medical Informatics, Epidemiology and Statistics Faculty of Medical Sciences, University of Nijmegen P.O. Box 9101, 6500 HB Nijmegen, email: p.krabbe@mie.kun.nl

"realistic" than the rating scale. As a fifth method we added paired comparisons, a common psychometric indirect scaling method. Paired comparisons is considered to be the best scaling method from a cognitive point of view. It is based on less complicated binary choices instead of the direct assessments that are required for the other four methods. Paired comparisons is based on measurement theory (Torgerson, 1958) and was used in one of the first studies which focused on the elicitation of valuations for health states (Fanshel and Bush, 1970).

Throughout the experiment, the EuroQol classification was used and all the design features of the EuroQol valuation questionnaire were applied, except those related to the valuation technique (EuroQol Group, 1990). The generic EuroQol descriptive system for health states is suitable for all valuation methods and has been used extensively in fundamental and applied valuation research (EuroQol Group, 1990; Essink-Bot *et al.*, 1993; Agt van *et al.*, 1994; O'Hanlon *et al.*, 1994; Selai and Rosser, 1995).

A few studies have focused on to the simultaneous comparison of more than two methods (Torrance, 1976; Bombardier *et al.*, 1982; Llewellyn-Thomas *et al.*, 1982; Sutherland *et al.*, 1983; Read *et al.*, 1984; Hornberger *et al.*, 1992; Bass *et al.*, 1994; O'Brien and Viramontes, 1994). Most studies only partially standardised the stimuli and the testing conditions, hampering the interpretation of interstudy differences and preventing replication. In the experimental study described here we have tried to pay close attention to differences caused by the methods themselves instead of unintentional local conditions.

Most of the theoretical assumptions underlying the current valuation methods, though tenable, have yet to be empirically proved and there is evidence that some of the assumptions need adjustments (Johannesson *et al.*, 1994; Verhoef *et al.*, 1994; Gafni, 1995; Wakker and Stiggelbout, 1995; Bleichrodt, 1996; Stalmeier *et al.*, 1996). However, the present study is not oriented towards the testing of the underlying assumptions of the five methods. This paper essentially focuses on two questions: (1) to what extent do the five valuation methods yield comparable results, and (2) which of the methods is statistically the most reliable?

The first question deals partially with the validity of the methods. Validity encompasses three main aspects each with a rather broad scope: content validity, criterion-related validity and construct validity. Content validity refers to the question: "Is the instrument really measuring what we intend to measure?" For the purpose of this study, this implies a discussion about the "real" meaning and interpretation of values elicited by valuation methods. Are they really representing individual expressions of health-state preferences? Criterion-related validity is only applicable if one method can be identified as superior, i.e. a "gold standard". As these issues are part of an ongoing debate (Froberg and Kane, 1989b; Nord, 1992), content and criterion-related validity were not investigated directly in this study. Here we are primarily dealing with convergent validity which may be regarded as a type of construct validity. Convergent validity was studied by examining equivalence and comparability. First, we investigated the equivalence of the valuation methods, e.g. are particular health states *absolutely* valued equally by different valuation methods? Second, we investigated comparability, a broader concept related to the *relative* relationship between valuation methods. Equivalence was tested by comparisons of raw values, comparability allows for (restricted) transformation of data (e.g. value functions).

As part of a recently proposed standard approach to the comparison of methods (Streiner and Norman, 1995), we studied the different sources of measurement error which enabled us to reveal the reliability of the valuation methods in detail.

MATERIAL AND METHODS

The health-state descriptions

For the description of health states we used the classification developed by the EuroQol Group (Brooks, 1996). The EuroQol classification describes health status according to five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each dimension has three levels, i.e. "no problems" (1), "some problems" (2), "severe problems" (3). "Holistic", "comprehensive" (Froberg and Kane, 1989a) health-state descriptions are composed by taking one level for each dimension (e.g. the best health state is represented by 11111). Theoretically this set of dimensions and levels of the EuroQol instrument allows for 243 (3^5) different health-state descriptions ("vignettes"). The EuroQol Group selected 11 of these vignettes as a standard set for experiment and study. Two health states were added in the present experiment as previous results indicated that the original set did not evenly cover the continuum between 0 (zero) and 100. Within each separate experiment, the vignettes to be valued were presented in a randomised order to avoid memory effects.

Short description of the five valuation methods

Rating scale (RS). The rating scale used in this study was the EuroQol "thermometer". This rating scale is presented as a vertical thermometer with a scale from 0 to 100. The anchors were labelled "best imaginable health state" at the top of the thermometer (100) and the "worst imaginable health state" at the bottom (0). The participants'

task was to locate 13 different vignettes on a scale between the two anchors to represent their preference, in such a way that the intervals between the positions of the vignettes corresponded to the differences in preference as perceived by the participant. The task required the respondents to locate all the vignettes on the same scale.* This scaling task differed from the other three direct valuation methods (standard gamble, time trade-off, willingness-to-pay), which required the health states to be valued separately.

Standard gamble (SG)

The standard gamble method comprises an iterative paired comparison. SG involves making choices under conditions of uncertainty. Participants have to reach a point of indifference between the two alternatives by varying one of them using a "ping-pong" strategy. Conventionally, SG is operationalised as a choice between being in a specific lifelong stationary impaired health state (the state to be valued) or a hypothetical procedure with two outcomes: a probability (p) of instantaneous and lasting improvement to perfect health; or a complementary probability ($1-p$) of, conventionally, immediate death. By varying the p-level of the uncertainty outcome, the point of indifference between the two alternatives is determined. The value (utility) of the stationary state is defined as the probability p at the indifference situation. SG is regarded as a valid operationalisation of the von Neuman–Morgenstern utility gamble (Neumann von and Morgenstern, 1953; Cohen, 1996).

The descriptions of the best outcome, generally described as "perfect health" or "optimal health" in other studies, were phrased as "best imaginable health state" in the present study. Our operationalisation of SG differed from most publications in the choice of the lower anchor point in the gamble. We replaced "being dead" with the "worst imaginable health state" primarily for reasons of standardisation between methods. This choice can be justified based on the assumptions of the method (Torrance, 1986; Llewellyn-Thomas *et al.*, 1982). It was clearly stressed to the participants that both outcomes arising from the gamble would involve chronic health states. Values obtained in this way require a linear rescaling factor to be comparable with values obtained with the standard SG, assuming perfect scalability of "dead" and the "worst imaginable health state" on the assumed health continuum (Krabbe *et al.*, 1996).

Time trade-off (TTO)

This method was developed by Torrance as a less complicated, conceptually different although equally sound alternative to SG. Like SG, time trade-off is based on trade-offs, but the concept of uncertainty is omitted. Participants trade-off survival time and health status. In the conventional operationalisation, the first alternative offers a (suboptimal) stationary health state with a given duration (x), 10 years in the present study. A better health status (conventionally perfect health) of shorter duration is offered as the competing alternative, conventionally followed by death. The point of indifference is reached by varying the duration spent in perfect health (y). Subsequently, by combining x and y, the value of the stationary health state can be established (y/x). For reasons already mentioned, we replaced "being dead" by "worst imaginable health state" in the present study (Krabbe *et al.*, 1996). The optimal health state was phrased as "best imaginable health state". For both options the health state would return to its present form after ten years.

Willingness-to-pay (WTP)

The willingness-to-pay task in our study started by confirming the average budget situation of the medical students participating in the experiment. A monthly budget of $725 (standard study grant of 1200 Dutch florins given by the government, 1993) could be spent after subtraction of the rent for a room and fixed costs for food, heating, clothing etc. (500 Dutch florins). Respondents were asked to imagine that they were in a certain impaired state of health and asked what amount they were willing to give up permanently to return to their previous (healthy) condition. This operationalisation was chosen after piloting available alternatives.

Paired comparisons (PC)

Paired comparisons is a scaling method consisting of a two-step procedure (McIver and Carmines, 1981). PC is especially developed for scaling unconcatenate subjective attributes (such as: food, politicians). As in the three trade-off methods, the participant is confronted with two outcomes, but here preference is required rather than trying to achieve a point of indifference. The data on individual preferences between all possible pairs of health-state descriptions allow for the construction of a matrix of $1/2(n(n-1))$ preferences, expressed as probabilities. The probability in every cell of the matrix is the proportion of the "row" health state being preferred to the "column" health state by the judging panel. As a second step, transformations and computations based on scaling theory, construct a unidimensional interval scale of health states.

*Usually the application of RS implies that for each stimulus valuation a separate rating scale is used. Here the health states were valued simultaneously in two sets of vignettes on facing pages, with on each page a vertical scale.

In this experiment, the 13 health states to be scaled, comprised a considerable number of paired comparisons consisting of so-called dominant pairs, i.e. one of the two health states is objectively "by definition" better than the other health state (e.g. "12232" is more severe than "12132"). Out of 78 possible pairs* 43 pairs were dominant. For reasons of efficiency only the remaining 35 non-dominant pairs (45% of all the possible pairs) were valued.

After the standard forced choice comparison, we requested a graded choice (scale 1...9: 1 = strong preference health state A, 5 = indifferent, 9 = strong preference health state B) (see also Hadorn *et al.*, 1992).

ORGANISATION AND TESTING CONDITIONS

The experiment included two sessions, separated by a 10-day interval. The same group of 104 students participated in both sessions. Students were recruited by handouts. For full participation they were paid a fee of approximately $65 (1993). Data collection took place in a group, since another objective of the experiment was to study the equivalence of collectively and individually collected responses (published elsewhere; Krabbe *et al.*, 1996). Both sessions consisted of a sequence of valuation tasks deliberately interspaced with unrelated questionnaires, e.g. on the moral acceptability of genetic manipulation. From pilot studies with other participants we learned that weariness and even irritation due to monotony had to be prevented by alternation of tasks and the inclusion of breaks.

All participants were seated in a lecture hall with due space between them. Each different method was preceded by a similar verbal explanation of the method and a few test judgements. The descriptions of the health states to be valued were always presented by slide projection. During the presentation, the instructors (GJB, MLE-B) repeated the nature of the particular method for each valuation to avoid blurring of the concepts.

Values for the methods RS and WTP were elicited during the first session, SG and TTO during the second session. The collection of the PC method responses were divided over the two sessions (both PC forms alternate for each health state).

Responses were collected by pencil and paper for RS, SG, TTO and WTP, and by means of an electronic response system (standard PC: choose A or B) for the two types of PC methods. For SG each participant responded by dividing a "probability pie" into two complementary parts. The task for TTO was to divide a "duration bar" into two parts.

All the separate tasks were pretested with other panels and a detailed work schedule was used to ascertain equivalence of presentation, of explanation, etc.

In order to detect differences associated with characteristics of the methods themselves, we controlled for the following:

1. factors related to the health states (such as prognosis) were kept constant;
2. factors related to the subjects who performed the valuation tasks (age, education, experience with illness) were kept constant by selecting a homogenous panel; and
3. characteristics of presentation of the health-state descriptions (order, framing, layout, instructions) by written protocols and training.

ANALYSIS

Outcomes of the RS, SG, TTO and WTP methods were transformed by linear transformation to a uniform 0–100 scale (RS = score; TTO = 10 × score; SG = 100 × score in degrees/360; WTP = [DFl.1200 – DFl.500 – score]/7).

To analyse the partial preference matrix of the PC task we used Thurstone scaling (Hadorn *et al.*, 1992; Torgerson, 1958) to derive a unidimensional, interval scale of health-state preferences. For the graded paired comparisons task, we computed the average preference rating (APR) as described by Hadorn *et al.* (1992). We included all the responses for each of the five methods, although they were not fully complete due to missing values (see below).

If a valuation method is cognitively easy to handle and clear to understand (feasibility), it might be expected that in dominant pairs of health states, the better state is preferred. If this is not the case the results are viewed as inconsistent. In order to study inconsistencies in the valuation of dominant pairs, we computed distances between the dominant health state and the secondary health state for all relevant pairs. According to our definition, the distance between vignettes "33332" and "11112" is the summation of the level differences for the five dimensions. For this example the distance is: (3 – 1 =)2 for dimension one-four and is (2 – 2 =)0 for the last dimension, yielding a total distance of 8. Vignette "33332" had the largest distance in relation to vignettes "11112", "11121", "11211", "12111", and "21111". Respectively the smallest distance was between vignette "11122" and two vignettes "11121" and "11112".

Validity—simple. Convergent validity between the methods, based on the mean values of the health states, was investigated by Pearson's product–moment correlation coefficient and Spearman's rank correlation coefficient. The first statistic is suitable for interval or ratio data while the second statistic is more appropriate for ordinal data or for

*$\frac{1}{2}(n(n-1)) = \frac{1}{2}(13 \times (13-1)) = 78$.

Table 1. Mean values and standard deviations for the 13 health states (n = 97–104; between parentheses SG order) by the four methods (all linearly transformed to 1–100)

	Standard gamble[a]		Time trade-off[b]		Rating scale[c]		Willingness-to-pay[d]	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
12111 (1)	96.2	5.3	94.4 (1)	8.8	68.2 (5)	12.8	89.5 (1)	9.1
11211 (2)	95.6	4.4	92.6 (4)	7.7	73.4 (1)	11.7	86.8 (3)	11.2
21111 (3)	94.5	7.0	92.8 (3)	8.7	71.7 (2)	10.1	83.1 (5)	14.6
11112 (4)	93.8	12.3	93.6 (2)	8.0	70.3 (3)	11.8	87.2 (2)	13.9
11121 (5)	93.3	8.9	91.8 (5)	8.6	70.2 (4)	12.7	86.3 (4)	14.6
11122 (6)	89.0	13.5	86.0 (6)	11.0	55.0 (6)	12.0	81.1 (6)	15.6
12212 (7)	81.7	15.1	78.6 (7)	14.4	47.0 (7)	12.7	75.9 (7)	13.9
32211 (8)	79.2	18.6	73.1 (8)	18.6	41.2 (8)	12.8	65.3 (8)	19.1
21232 (9)	65.2	22.8	59.0 (10)	20.3	31.1 (9)	14.2	60.1 (9)	19.2
22323 (10)	64.5	23.7	61.0 (9)	22.6	24.6 (11)	13.0	59.5 (10)	17.5
33321 (11)	53.6	26.5	47.8 (11)	24.0	26.4 (10)	12.7	58.6 (11)	18.4
22233 (12)	51.5	28.4	44.9 (12)	24.7	22.1 (12)	13.4	52.0 (12)	18.4
33332 (13)	34.4	25.3	27.8 (13)	23.4	10.7 (13)	9.0	45.6 (13)	18.9
Range mean	61.8		66.6		62.7		43.9	
Mean SD		16.3		15.5		12.2		15.7

[a]SG; scores transformed as: SG = 100 × score in degrees/360;
[b]TTO; scores transformed as: TTO = 10 × score;
[c]RS; untransformed scores;
[d]WTP; scores transformed; recoded as: WTP = (700 − WTP original)/7.

data of higher measurement level that does not satisfy requirements for r. To test exact concordance of continuous data, we also computed intraclass correlation coefficients (ICCs). ICCs include level effects between different measurements. These three coefficients for convergent validity were all computed based on the mean values for the 13 health states.

Validity—construct. To study construct validity for the four direct valuation methods (PC could not be included being an indirect scaling method, yielding a different type of data) we applied the multitrait–multimethod methodology (MTMM) on the individual responses (Crocker and Algina, 1986; Hadorn and Hays, 1991). Based on a matrix representing all the intercorrelations between multiple traits (13 health states) and multiple methods (RS, SG, TTO, WTP), four classes of correlations can be distinguished (see Appendix A).

Validity—convertability. We examined the numerical comparability among the methods. If valuation methods are not equivalent (i.e. they do not give the same values for instance, intraclass correlations coefficients are not high), perhaps values are related in some systematic way so that conversion curves can be constructed. Power functions [method $Y = 1 - (1 - \text{method } X)^{\alpha}$], similar to Torrance (Torrance, 1976), were therefore estimated relating mean values of the 13 states for all six pairs of methods. Computations have been performed by the non-linear regression module of SPSS for Windows.

Measurement error/reliability. We used Generalisability Theory (G-theory) as a general approach to estimate the relative contribution of the multiple sources (facets in G-theory language) to measurement error/bias (Streiner and Norman, 1995; Krabbe *et al.*, 1996). G-theory is a specific application of analysis of variance (ANOVA) and requires individual data. In the present case, the relative contribution (variance components) of the facets "health state", "method" and "participants", their interaction terms as well as all other facets of measurement error, were estimated separately.

Furthermore, G-theory was used to estimate reliability coefficients for the separate methods. These reliabilities are closely related to the internal consistency concept (Cronbach's alpha). Although instead of the stimuli (health states), here the responses of the participants to the stimuli as they were elicited by the valuation methods were tested.* Hence, agreement among participants was estimated rather than similarity of items (health states). G-theory is a method which treats valuations at interval measurement level. In order to study the internal consistency reliability among the participants in their valuation of the set of 13 health states, but treating the valuations as rankings, Kendall coefficients of concordance W were determined (Siegel and Castellan, 1988), concurrent with the G-study.

RESULTS

Response

Of the 104 participants in this study, 46% were male. All were students, 71% were medical students. Mean age of the group was 22 (SD = 2.48) years. RS, SG, TTO and WTP each took about 15 minutes to complete. The responses of all 70 PC

*The reason for this exchange is that we are dealing with data that stems from a so-called stimulus-scaling task (see Froberg and Kane, 1989b).

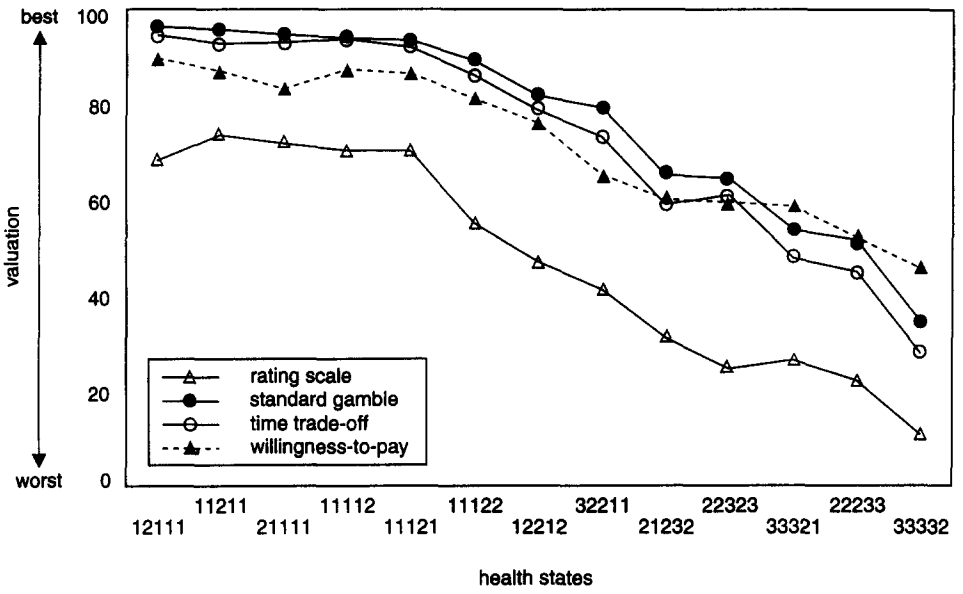

Fig. 1. Valuations (means) for the 13 EuroQol health-state descriptions elicited by the four methods (ordered by standard gamble values).

pairs (35 standard and 35 graded) using the voting system took about 90 minutes. Judging from participants' remarks and from the absence of learning effects, we regarded memory effects to be highly unlikely. The feasibility of these experiments was satisfactory, although at the end some participants complained of weariness. Few responses were missing.

Descriptive statistics and consistency measurement

The results of the paired comparisons method proved to be unstable. If the empirical data were changed in only a minor way this resulted in a major alteration of the unidimensional scale. The background to this was the decision to leave out the apparently dominant pairs from the empirical task, leaving the relatively difficult ones to be measured and scaled. Empty cells were substituted with "expected" preferences, but this approach also yielded unstable results and was therefore rejected. Thus no results from the PC method will be presented in this paper.

Table 1 shows the results (means and standard deviations for each health state and overall means and ranges for the methods) of the experiment for the four remaining methods. The order of presentation of the 13 health states in Table 1 is arbitrarily based on the SG values. Mean valuations for the 13 health states for the four methods are shown in Fig. 1.

A summary of the measures for inconsistency (individual level) is presented in Table 2. As expected, the inconsistency was highest for the pairs with the smallest distances. Average inconsistency for the methods SG and TTO was almost the same (4.6% and 4.3%, respectively). In RS it was lower (2.0%) and in WTP higher: 7.4%. We observed a 50.5% inconsistency for the method WTP for the two dominant pairs with distance 2.

Validity

Figure 2 shows the correlations between the four methods as a first estimate of convergent validity. The Pearson product–moment correlation coefficients were high and close to 1.0 for all the six relationships between the four methods. Spearman rank correlations were slightly lower than the interval-based Pearson product–moment correlation coefficients. In Fig. 1 the four lines do not match but are parallel. ICCs were much lower than the Pearson and Spearman correlations, particularly for the relationship between RS and the other methods, suggesting important level effects.

Table 2. Percentages of inconsistencies between dominant pairs[a] of health states for each of the four methods[b] (n = 97–104)

Distance	Number of pairs	RS	SG	TTO	WTP
1	2	12.5	21.6	17.8	50.5
2	5	10.6	13.7	12.1	21.9
3	5	6.2	8.5	6.3	11.5
4	7	3.2	7.1	5.1	7.8
5	3	2.6	4.8	5.4	4.5
6	15	0.4	2.9	3.0	6.3
7	1	1.0	0.0	1.9	1.9
8	5	0.0	1.9	2.3	2.9
Total	43	3.4	6.3	5.5	10.6
Weighted total[c]		2.0	4.6	4.3	7.4

[a]Total number of valuable dominant pairs: 43 × 104 = 4472;
[b]RS = rating scale, SG = standard gamble, TTO = time trade-off, WTP = willingness-to-pay;
[c] weighted total: $\frac{\text{SUM(number of pairs} \times \text{proportion of inconsistencies} \times N)}{\text{SUM(distance} \times \text{weighted number of pairs)}}$.

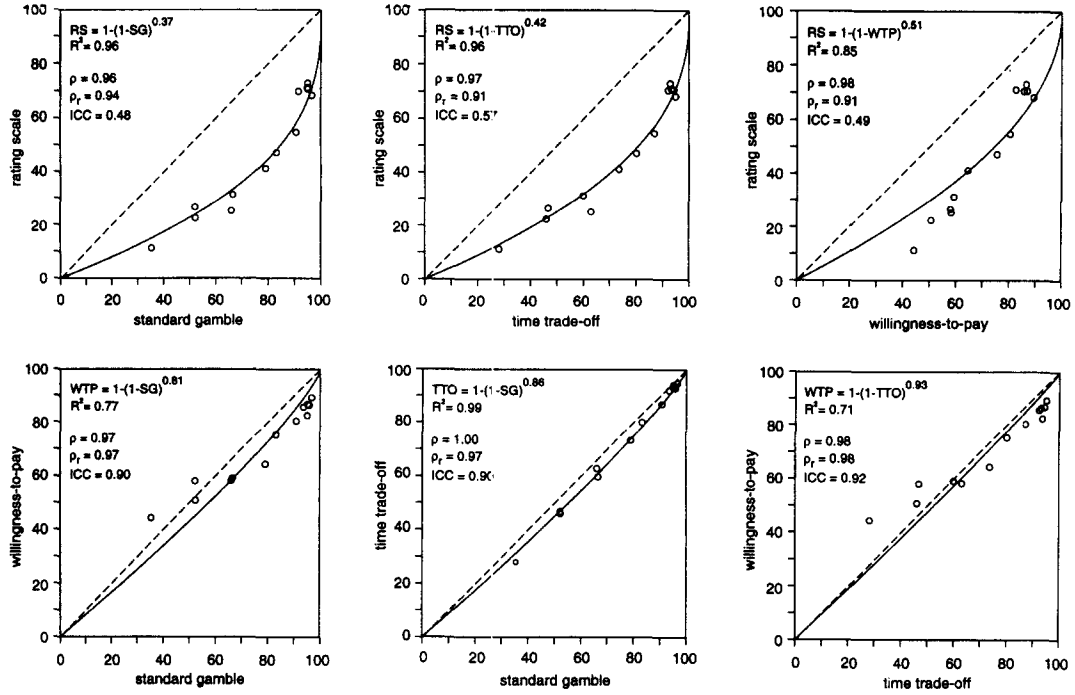

Fig. 2. Convergent validity measured by correlation measures (ρ = product–moment Pearson correlation, ρ_r = Spearman rank correlation, ICC = interclass correlation) between the health-state values for the four methods based on the mean values (means of Table 1, n = 13) and functional relationships between the four elicitation methods studied by non-linear regression analysis (power function); the entry labelled R^2 is the coefficient of determination and may be interpreted as the proportion of the total variation of the dependent variable around its mean that is explained by the fitting model. .

Due to specific patterns (see Conclusions and discussion) between the correlation coefficients of the data computed for the MTMM methodology, only a partial analysis was valid. This is presented in Table 3 which shows that convergent validity (monotrait–heteromethod coefficients)* was reasonably good for all the health states across the methods SG and TTO (overall: 0.69). All other comparisons between the four methods showed low convergent validity (overall: 0.15 – 0.25). Coefficients for the comparisons between trade-off pairs WTP/SG and WTP/TTO were even lower than for RS/SG and RS/TTO. No particular pattern could be detected by inspection of the means of the convergent validity coefficients of the 13 health states in Table 3.

Except for WTP, values of all the other methods appeared to be exchangeable after a one-parametrical (α) power transformation. The percentage of variance explained by the power functions was: RS = f(SG), 96%; RS = f(TTO), 96%; TTO = f(SG), 99%. Accuracy of predictions including WTP was considerably lower: RS = f(WTP), 85%; WTP = f(SG), 77%; WTP = f(TTO), 71%. All plots of RS with the trade-offs showed concave power functions (α = 0.37–0.51), the other functions were approximately linear (a = 0.81–0.93).

*Because these correlation coefficients are based on the analysis of individual values for the 13 health states of the four methods, such correlations are therefore always lower than (Pearson PM) correlation coefficients based on aggregated mean data (e.g. Fig. 2).

Measurement error and reliability

The initial analysis on the individual raw scores demonstrated that only 44.8% of the total variance was attributed to the variability of the 13 health states (H) scores (see Table 4). A small percentage of variance, 4.7%, was accounted for by systematic differences in valuations of the health states by the facet participants (P). This relatively small contribution indicated that, averaged over all the health states and all the methods, the participants valued the health states only slightly systematically differently. Twenty percent of the total variance stemmed from the facet methods (M), which was in particular attributable to the divergent magnitude of the RS valuations. Overall, 17.4% of the total variance was attributable to the first-order interaction terms (HP, HM, PM). The interpretation of, for instance, the interaction term HP is that *some* participants valued *some* health states systematically differently. Only 13% of the total variance came from the three-way interaction HMP, which suggests a satisfactory explanation model for this data set.

Additionally the individual values for the health states were transformed to method-specific z-values. Absolute differences between the four methods (M), especially between the RS and the other methods,

Table 3. Convergent validity (monotrait–heteromethod validity correlations = same health state assessed by different methods) for the 13 health states and the four methods[a] based on the individual values (n = 91)

	12111	11211	21111	11112	11121	11122	12212	32211	21232	22323	33321	22233	33332	Mean methods[b]
Comparison														
RS vs SG	0.46	−0.03	0.08	0.27	0.20	0.28	0.07	0.14	0.12	0.25	0.20	0.22	0.17	0.22
RS vs TTO	0.36	0.07	0.22	0.19	0.20	0.33	0.20	0.19	0.18	0.33	0.20	0.20	0.10	0.23
RS vs WTP	0.10	0.32	0.26	0.29	0.23	0.40	0.19	0.24	0.33	0.07	0.21	0.13	0.19	0.25
SG vs TTO	0.63	0.66	0.72	0.59	0.84	0.77	0.52	0.65	0.67	0.74	0.73	0.70	0.67	0.69
WTP vs SG	0.16	0.14	0.14	−0.02	0.29	0.16	0.12	0.13	0.07	−0.03	0.17	0.15	0.10	0.15
WTP vs TTO	0.25	0.18	0.19	0.19	0.24	0.25	0.09	0.21	0.17	−0.04	0.12	0.11	0.07	0.18
Mean correlations between the four methods per health state[b]	0.37	0.32	0.34	0.31	0.40	0.41	0.25	0.32	0.33	0.35	0.34	0.32	0.30	

[a]RS = rating scale, SG = standard gamble, TTO = time trade-off, WTP = willingness-to-pay;
[b]For each health state and for each comparison between two methods the square root of the mean of the squared correlations was computed to summarise the rows/columns with correlation coefficients.

were eliminated by this standardisation, yielding by definition zero variance for the factor methods (M) and increasing the percentage of variance uniquely attributable to the health states (H) to 72%.

The results of the G-study for each of the methods separately are shown in Table 5. For the WTP method more than 30% of the total variance was due to systematic differences between participants valuing the 13 health states, which was high compared to the other methods. Additional inspection of the data revealed that the relatively great contribution of systematic differences for WTP between the participants was reducible to two response patterns. One response pattern consisted of a small trade-off of the budget except for the very severe health states, for which almost the whole budget was exchanged to remain in full health. The other response pattern showed exchange of almost the whole budget for even moderately bad health states (insensitivity for the stimulus, e.g. due to cognitive difficulty).

The variance components of the health states (H) can be regarded as (standard) reliability coefficients, assuming interval metric properties of the data. RS was the method with the highest reliability: 0.77. The reliability for WTP, 0.49 was low.

Non-parametric statistics revealed higher coefficients. For all the four methods there was good agreement among participants in their ranking of the health states (Kendall's W-test; n = 91). The highest agreement, 0.83, was achieved by the RS method. Agreement using SG was sufficient: 0.75. For TTO and WTP the coefficients were 0.77 and 0.80, respectively.

CONCLUSIONS AND DISCUSSION

Under highly controlled conditions we conducted an experiment with five valuation methods. Design characteristics aimed at maximal standardisation except for the two manipulated effects, i.e. health states and valuation methods. We assumed these to be responsible for the observed effects. Although we were able to control for many factors, other factors may still have influenced the results.

In this study, session effects are the most conceivable ones. Generalisation of the study results may be further restricted due to the composition of the panel that performed the valuation tasks and to the selection of the 13 health states. A different selection of respondents could lead to different results, although several studies have shown that these effects in this context are minor (Essink-Bot *et al.*, 1993; EuroQol Group, 1990). The selection of the health states may have to some extent influenced the results of this study, although we expect the relationship between the methods to be hardly influenced because the sample of the health states was well chosen.*

Table 4. Estimated variance components (percentages) of health states (13) × participants (91) × methods (4)

	Variance components	
Source of variation	Raw scores	Individual scores transformed for each method to z-values
Health states (H)	44.8	72.0
Methods (M)	20.0	–[a]
Participants (P)	4.7	0.7
HM	2.3	2.0
HP	8.0	8.8
PM	7.1	1.2
Residual (HMP, e)	13.1	15.3

[a]By definition.

*An alternative study based on a set of 13 health states selected by a restriced inclusion criterium (for example, 13 EuroQol health states, without level 3) would induce a decline of the proportion of variance (Tables 4 and 5) for the facet "health states" and consequently yield lower reliability coefficients. However, we were not interested in the characteristics of valuation methods for a specific domain of health states. The selection of the health states was deliberately worked out to evenly cover the continuum between 0 and 100, which enables us to study the "behaviour" of the participants for the valuation methods on the whole range of possible health states. In this context we are particularly interested in the comparison between the methods. The inclusion of a set of health states with a restricted range would have certainly decreased the proportion of variance contributed to the health states, but would also obscure the division between the methods.

Table 5. Estimated variance components (percentages) of the health states (13) × participants (91) for each method[a] separately

Source of variation	Variance components			
	RS	SG	TTO	WTP
Health states (H)	77.0	57.6	64.6	48.9
Participants (P)	5.5	11.8	9.9	31.4
Residual (HP, *e*)	17.6	30.6	25.5	19.7

[a] RS = rating scale, SG = standard gamble, TTO = time trade-off, WTP = willingness-to-pay.

We will first clarify the unexpected outcomes of the WTP method and the problems that we confronted using the PC scaling method. We will then discuss the comparability of the methods and their reliabilities. Finally we will consider the complications we encountered when studying construct validity using the MTMM methodology.

We found it difficult to proceed with the WTP method, even among this homogeneous and highly educated population and despite our controlled study design with extensive explanation and test questions. Two typical response patterns appeared to determine the reliability and the range of the responses. Thus the WTP results were not satisfactorily comparable to the other trade-off methods despite satisfactory regression results and inter-method comparability on first sight. Although a linear transformation of the mean WTP values to SG/TTO was technically possible, WTP in our operationalisation was found to be an inferior method with an unacceptably low reliability. Even more worrying was the amount of inconsistency found between the dominant pairs of health states. Evidence from the few studies that have focused on WTP is difficult to interpret due to variability of concepts used, the small samples, and the small number of health states which do not allow for sound statistical testing (Thompson *et al.*, 1982, 1984; Thompson, 1986; O'Brien and Viramontes, 1994; Chestnut *et al.*, 1996). Unless it is possible to improve the operationalisation of WTP it may have to be regarded as an unfeasible method. Perhaps therefore, the concept of WTP is only valid in real life situations (sometimes called revealed preference or averting behaviour method) and not suitable under experimental conditions.

Serious problems were also encountered with the PC scaling method as it did not provide stable outcomes for both the PC variants (standard, graded). The underlying difficulty with the application of PC is the high number of pairs to be valued with a complete design and the probability of bias in dominant pair evaluation. Due to the partially ordered nature of our stimuli, we could not overcome the problem with the relatively high number of empty cells (dominant pairs) (MacCallum, 1978).

Hadorn *et al.* (1992) applied PC with a partial design with apparently more success ($n = 93$). In our analysis the factor critical to failure appeared to be the number of dominant pairs and the level of complexity of the classification. Hadorn *et al.* used only two dimensions (i.e. "pain or physical suffering" and "limitations on daily activities") with four levels each (EuroQol: five dimensions, three levels) and only selected relatively comparable pairs of health states. Therefore their PC analysis was based on an incomplete and selective design of 54 (59%) of the total number of pairs.* Reconstruction of Hadorn's design revealed that still 40% of these 54 pairs were dominant pairs (in our design 0%). Moreover, no mention was made by Hadorn *et al.* of the stability of the PC method for scaling of health states based on their incomplete design nor did they report the effect of the substitution of empty cells with "expected" preferences.

A surprising finding was the performance of the MTMM methodology in this context. The method was advocated by Froberg and Kane (1989c) for good reasons and empirically applied by Hadorn and Hays (1991). In retrospect, our failure with MTMM can be explained by the characteristics of data yielded by the process of valuing "subjective" stimuli such as health states as opposed to the more common situation where participants have to reveal their opinion on, for instance, the attractiveness of consumer goods with "latent" indivisible characteristics. In our study, health states have "manifest", ordered domains. If dominance exists, as is the case here, then the usual MTMM analyses are not adequate. Correlations between the health states then show a special structure indicated as a "simplex structure" (Jöreskog and Sörbom, 1979). The typical property of a simplex correlation matrix is that correlations decrease as one moves away from the main diagonal. Valuations of health states that were of the same severity will show moderate between-method correlation, but valuations of health states that were different in severity (e.g. "21111" vs "33321") show no correlation at all (as was observed in our data). MTMM analysis requires at least moderate or low correlations among all health states elicited by one and the same method.

Hadorn and Hays (1991) presented an early application of MTMM analyses. Six aspects of health-related quality of life (HRQOL) were investigated (i.e. general health perception, physical suffering etc.). Participants ($n = 76$) were asked to provide preference ratings (valuations) by judging the effects of different levels of problems or impairments on

*To the initial 16 constructed health states they excluded the two anchor states (no suffering—no limits, severe suffering—severely limited), which resulted into $1/2(14 \times (14-1)) = 91$ pairs of vignettes to be compared. The authors reported that despite the dominance restriction, 54 pairs remained to be assessed (after reconstruction of their design we arrived at 50 pairs). Therefore, at least 37 pairs were not valued as it was thought that one of the health states of such a pair was manifestly dominant.

each of the six dimensions on overall quality of life. This task was performed for three different assessment methods, developed by the authors themselves. As a result of their different strategy which was not dealing with the valuation of health states but with eliciting individual preferences for *separate* aspects of HRQOL, the problem of the simplex structure that we encountered with MTMM was absent. After some consideration we judge MTMM incompatible with data analysis of standard n (independent domains) $\times p$ (ordered levels) classification systems.*

We investigated the convertibility of the methods straightforwardly applying simple algebraic power functions. Torrance (1976) reported a power relationship $RS = 1-(1-TTO)^{\alpha}$ between RS and TTO with a coefficient of 0.62 ($R^2 = 0.80$) based on 18 means of valued health-state scenarios (n = approx. 200). In a study by Stiggelbout *et al.* (1996) a coefficient of 0.64 was presented. Loomes (1993) found a coefficient of 0.55 based on a secondary analysis of data by Bombardier *et al.* (1982). We found, based on 13 mean values, $\alpha = 0.42$ ($R^2 = 0.96$) for the power function. Busschbach (personal communication, 1996) reported similar results, namely $\alpha = 0.47$ ($R^2 = 0.95$; $n = 103$). Different coefficient values may be the result of many factors. Of the 18 scenarios in Torrance's study, none were valued very low or high, which may have caused the higher power coefficient. The study of Stiggelbout *et al.*, even more than Torrance's study, lacked a broad range of health states because each respondent valued only his/her own health state. Other factors that could be responsible for different outcomes are: the composition of the valuation panels, the instruction to the panel and the classification system used.

We conclude that valuations of health states based on rating scales are distinct from but strongly related to outcomes derived through trade-off methods. Trade-off methods elicit values expressing an individual's preference for a particular health state under a condition where something has to be *sacrificed* (e.g. change on good outcome, life-years, budget). Rating scale methods however are based on the *comparison* of different health states. RS values express the subjects' internal representation of health states in a stable world where the actual health of the respondent probably plays a major role as a reference point.

The choice of which type of values is to be used depends largely on the perspective of application. From the individual perspective, generally directed at decisions on change, trade-offs seem more appropriate to elicit valuations. For collecting societally grounded health-state valuations the RS method presumably is a feasible tool, particular if ordering of health states is the restricted goal. Use in the context of societal decisions theoretically requires power transformation.

The reliability coefficients estimated by the G-study showed lower reliability for all the methods in comparison with Kendall's W concordance coefficient based on ranks. Reliability of a G-study takes not only the ordering of health states into account but also the distances between health-state values. This explains why, in the case of WTP, the G-theory reliability coefficient was only 0.49 vs the Kendall's W of 0.80. Reliability was satisfactory for SG and TTO. In this study the RS method showed a reliability even higher than the two standard methods (see also Torrance, 1986).

Taken together, a valid comparison of more than two valuation methods under highly controlled conditions is feasible and a simple power transformation suffices to describe the value function between health-state valuation methods. The RS method is in this sense almost congruent to SG and TTO.

Two interesting "negative" outcomes require further study. First, the PC method proved to be not applicable due to the dominant pairs of health states. Valuations of only non-dominant pairs of health states impairs accurate estimation of scale values. Inclusion of all pairs of health states yields highly flawed results. Also, the MTMM methodology appeared not to be suitable for essentially the same reason as the failure of the PC method.

Future consideration might be given to whether there are other techniques/methodologies that are potentially valuable for the elicitation of valuations/preferences for health states. Unfolding analysis could be such a technique (Coombs, 1950; Lewis-Beck, 1995). It is fully focused on the analysis of preference data. Additionally, a methodology used with good results in a small number of fields is functional measurement (Anderson, 1976) and conjoint analysis (Louviere, 1988). A specific example of its implementation is the multi-attribute application of Torrance *et al.* (1982). But most of all well structured experiments and studies are needed to clarify the numerous indistinct concepts and assumptions related to the use of health-state valuation methods.

Acknowledgements—The authors wish to thank the Advisory Board for the Health Research Promotion Program (Adviesgroep SGO) for financially supporting the Research Program "Standardisation in Medical Technology Assessment". We would also like to thank C. Vonk, MSc. for providing the formula of the numbers of dominant pairs.

*The following two formulas relate to this topic. The number of pairs that can be achieved based on a descriptive system of p levels and n domains is $\frac{p^n(p^n-1)}{2}$. For computing the number of dominant pairs the formula is $\left[\frac{p^n(p-1)}{2}\right]^n - p^n$.

REFERENCES

Agt van, H. M. E., Essink-Bot, M., Krabbe, P. F. M. and Bonsel, G. J. (1994) Test–retest reliability of health state valuations collected with the EuroQol questionnaire. *Social Science & Medicine* **11,** 1537–1544.

Anderson, N. H. (1976) How functional measurement can yield validated interval scales of mental quantities. *Journal of Applied Psychology* **61,** 677–692.

Bass, E. B., Steinberg, E. P., Pitt, H. A., Griffiths, R. I., Lillemoe, K. D., Saba, G. P. and Johns, C. (1994) Comparison of the rating scale and the standard gamble in measuring patient preferences for outcomes of gallstone disease. *Medical Decision Making* **14,** 307–314.

Bleichrodt, H. (1996) Applications of utility theory in the economic evaluation of health care. Thesis, Erasmus University, Rotterdam.

Bombardier, C., Wolfson, A. D., Sinclair, A. J. and McGreer, A. (1982) Comparison of three preference measurement methodologies in the evaluation of a functional index. In *Choices in Health Care: Decision Making and Evaluation of Effectiveness*, eds. R. Deber and G. Thompson. University of Toronto, Toronto.

Brooks, R. (1995) *Health Status Measurement: A Perspective on Change*. Macmillan, London.

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

Campbell, D. T. and Fiske, D. W. (1959) Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin* **56,** 81–105.

Chestnut, L. G., Keller, L. R., Lambert, W. E. and Rowe, R. D. (1996) Measuring heart patient' willingness to pay for changes in angina symptoms. *Medical Decision Making* **16,** 65–77.

Cohen, B. J. (1996) Is expected utility theory normative for medical decision making? *Medical Decision Making* **16,** 1–6.

Coombs, C. H. (1950) Psychological scaling without a unit of measurement. *Psychological Review* **57,** 145–158.

Crocker, L. and Algina, J. (1986) *Introduction to Classical and Modern Test Theory*. Holt, Rinehart and Winston, New York.

Essink-Bot, M. L. (1995) Health status as a measure of outcome of disease and treatment. Thesis, Erasmus University, Rotterdam.

Essink-Bot, M. L., Stouthard, M. E. A. S. and Bonsel, G. S. J. (1993) Generalizability of valuation on health state collected with the EuroQol-questionnaire. *Health Economics* **2,** 237–246.

EuroQol Group (1990) EuroQol—a new facility for the measurement of health-related quality of life. *Health Policy* **16,** 199–208.

Fanshel, S. and Bush, J. W. (1970) A health-status index and its application to health-services outcomes. *Operations Research* **18,** 1021–1066.

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

Froberg, D. G. and Kane, R. L. (1989b) Methodology for measuring health-state preferences. II: scaling methods. *Journal of Clinical Epidemiology* **42,** 459–471.

Froberg, D. G. and Kane, R. L. (1989c) methodology for measuring health-state preferences. IV: progress and a research agenda. *Journal of Clinical Epidemiology* **42,** 675–685.

Gafni, A. G. (1991) Willingness-to-pay as a measure of benefits: relevant questions in the context of public decisionmaking about health care programs. *Medical Care* **29,** 1246–1252.

Gafni, A. G. (1995) Time in health: can we measure individuals' "pure time preferences"? *Medical Decision Making* **15,** 31–37.

Hadorn, D. C. and Hays, R. D. (1991) Multitrait–multimethod analysis of health-related quality-of-life measures. *Medical Care* **29,** 829–840.

Hadorn, D. C., Hays, R. D., Uebersax, J. and Hauber, T. (1992) Improving task comprehension in the measurement of health state preferences: a trial of informational cartoon figures and a paired-comparison task. *Journal of Clinical Epidemiology* **45,** 233–243.

Hornberger, J. C., Redelmeier, D. A. and Petersen, J. (1992) Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis. *Journal of Clinical Epidemiology* **45,** 505–512.

Johannesson, M., Pliskin, J. S. and Weinstein, M. C. (1994) A note on QALYs, time tradeoff, and discounting. *Medical Decision Making* **14,** 188–193.

Jöreskog, K. G. and Sörbom, D. (1979) *Advanges in Factor Analysis and Structural Equation Models*. Abt Books, Cambridge, MA.

Jöreskog, K. G. and Sörbom, D. (1989) *LISREL 7: A Guide to the Program and Applications*. SPSS, Chicago.

Krabbe, P. F. M., Essink-Bot, M. L. and Bonsel, G. J. (1996) On the equivalence of collectively and individually collected responses: standard gamble and time tradeoff judgements of health states. *Medical Decision Making* **16,** 120–132.

Lewis-Beck, M. S. (1995) *Basic Measurement*. Sage, London.

Llewellyn-Thomas, H., Sutherland, H. J. and Tibshirani, R. (1982) The measurement of patients' values in medicine. *Medical Decision Making* **2,** 449–462.

Loomes, G. (1993) Disparities between health state measures: is there a rational explanation? In *The Economics of Rationality*, ed. B. Gerrard. Routledge. London.

Louviere, J. J. (1988) Conjoint analysis modelling of stated preferences: a review of theory, methods, recent developments and external validity. *Journal of Transport Economics and Policy* **22,** 93–119.

MacCallum, R. C. (1978) Recovery of structure in incomplete data by ALSCAL. *Psychometrika* **44,** 69–74.

McIver, J. P. and Carmines, E. G. (1981) *Unidimensional Scaling*. Sage, Beverly Hills and London.

Neumann, J. vonn and Morgenstern, O. (1953) *Theory of Games and Economic Behavior*. Princeton University Press, Princeton.

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

O'Brien, B. and Gafni, A. (1996) When do the "dollars" make sense? Toward a conceptual framework for contingent valuation studies in health care. *Medical Decision Making* **16,** 288–299.

O'Brien, B. and Viramontes, J. L. (1994) Willingness to pay: a valid and reliable measure of health state preference? *Medical Decision Making* **14,** 289–297.

O'Hanlon, M., Fox-Rushby, J. and Buxton, M. J. (1994) A qualitative and quantitative comparison of the EuroQol and time trade-off techniques. *International Journal of Health Sciences* **5,** 85–97.

Read, J. L., Quinn, R. J., Berwick, D. M., Fineberg, H. V. and Weinstein, M. C. (1984) Preferences for health outcomes: comparison of assessment methods. *Medical Decision Making* **4,** 315–329.

Schmitt, N. and Stults, D. M. (1986) Methodology review: analysis of multitrait–multimethod matrices. *Applied Psychological Measurement* **10,** 1–22.

Selai, C. and Rosser, R. (1995) Eliciting EuroQol descriptive data and utility scale values from inpatients. *PharmacoEconomics* **8,** 147–158.

Siegel, S. and Castellan, N. J. (1988) *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York.

Stalmeier, P. F. M., Bezembinder, T. G. G. and Unic, I. J. (1996) Proportional heuristics in time tradeoff and

conjoint measurement. *Medical Decision Making* **16,** 36–44.

Stiggelbout, A. M., Eijkemans, M. J. C., Kiebert, G. M., Kievit, J., Leer, J. W. H. and De Haes, H. J. C. J. M. (1996) The "utility" of the visual anolog scale in medical decision making and technology assessment: is it an alternative to the time trade-off?. *Journal of Technology Assessment in Health Care* **12,** 291–298.

Streiner, D. L. and Norman, G. R. (1995) *Health Measurement Scales: A Practical Guide to their Development and Use*. Oxford University Press, Oxford.

Sutherland, H. J., Dunn, V. and Boyd, N. F. (1983) Measurement of values for states of health with linear analog scales. *Medical Decision Making* **3,** 477–487.

Thompson, M. S. (1986) Willingness to pay and accept risks to cure chronic disease. *American Journal of Public Health* **76,** 392–396.

Thompson, M. S., Read, J. L. and Liang, M. (1982) Willingness-to-pay concepts for societal decisions in health. In *Values and Long-term Care*, eds. R. Kane and R. Kane, pp. 103–125. Heath, Lexington.

Thompson, M. S., Read, J. L. and Liang, M. (1984) Feasibility of willingness-to-pay measurement in chronic arthritis. *Medical Decision Making* **4,** 195–215.

Torgerson, W. S. (1958) *Theory and Methods of Scaling*. Wiley, New York.

Torrance, G. W. (1976) Social preferences for health states: an empirical evaluation of three measurement techniques. *Socio-Economic Planning Sciences* **10,** 129–136.

Torrance, G. W. (1986) Measurement of health state utilities for economic appraisal: a review. *Journal of Health Economics* **5,** 1–30.

Torrance, G. W., Boyle, M. H. and Horwood, S. P. (1982) Application of multi-attribute utility theory to measure social preferences for health states. *Operational Research* **30,** 1043–1069.

Verhoef, L. C. G., de Haan, A. F. J. and van Daal, W. A. J. (1994) Risk attitude in gambles with years of life: empirical support for prospect theory. *Medical Decision Making* **14,** 194–200.

Wakker, P. and Stiggelbout, A. (1995) Explaining distortions in utility elicitation through the rank-dependent model for risky choices. *Medical Decision Making* **15,** 180–186.

APPENDIX A

Multitrait–multimethod Methodology

Originally this methodology was introduced by Campbell and Fiske (1959). They identified four classes of correlation coefficients. Firstly, monotrait–monomethod reliability correlations (health states measured twice for each method separately: test–retest). Secondly, heterotrait–monomethod correlations (different health states for the same method). Thirdly, heterotrait–heteromethod correlations (different health states assessed by different methods). Finally, monotrait–heteromethod validity correlations (same health state assessed by different methods). Using MTMM, construct validity is supported if correlations among different methods are high for a single trait (convergent validity), but correlations between the same methods measuring different traits are low (discriminant validity). Although Campbell and Fiske recommended visual inspection of the MTMM matrix for assessment of construct validity, recent additional modelling procedures (e.g. confirmatory factor analysis) have been developed which may lead to more unequivocal interpretation of such data (Schmitt and Stults, 1986; Jöreskog and Sörbom, 1989). We have performed analyses based on both classical Campbell and Fiske criteria and by using confirmatory factor analysis.

For the basic MTMM model based on confirmatory factor analysis, we treated each of the 13 health states as separate traits and the four valuation methods as separate methods. Another model was estimated by constructing three clusters of health states (mild, moderate, severe) as three separate traits. Models were also estimated assuming dependency (correlation) between the methods. For all models, various transformations (logit, arcsine, rescaling) of the data were used. However, none of these models led to meaningful outcomes as a consequence of the specific structure of the correlation matrix (see Conclusions and discussion).

All data available from the authors.