

Comparison of health state utility values derived using time trade-off, rank and discrete choice data anchored on the full health-dead scale

John Brazier · Donna Rowen · Yaling Yang · Aki Tsuchiya

Received: 8 February 2010 / Accepted: 5 September 2011 / Published online: 30 September 2011
© Springer-Verlag 2011

Abstract Recent years have seen increasing interest in the use of ordinal methods to elicit health state utility values as an alternative to conventional methods such as standard gamble and time trade-off (TTO). However, in order to use these ordinal methods to produce health state values for use in cost-effectiveness analysis using cost per quality adjusted life year (QALY) analysis, these values must be anchored on the full health-dead scale. The paper reports on two feasibility studies that use two approaches to anchor health state utility values derived from discrete choice data on the full health-dead scale: normalising using (1) the TTO value of the worst state and (2) the coefficient on the ‘dead’ dummy variable. Health state utility values obtained using rank and discrete choice data are compared to more commonly used TTO utility values for two condition-specific preference-based measures; asthma and overactive bladder. Ordinal methods were found to offer a promising alternative to conventional cardinal methods of standard gamble and TTO. There remains a large and important research agenda to address.

Keywords Ranking · Discrete choice experiment · Health state utility values · Preference-based measures · QALYs

JEL Classification I10 - health, General

Introduction

The status of preference-based measures of health for generating quality adjusted life years (QALYs) was considerably enhanced by the recommendations of the U.S. Public Health Service Panel on Cost-Effectiveness in Health and Medicine to use them in economic evaluation [1]. The use of preference-based measures has grown considerably over the last decade with the increasing use of economic evaluation to inform health policy, for example through the establishment of bodies such as the National Institute of Health and Clinical Excellence in England and Wales [2].

To be a preference-based measure, it has been suggested that the health state valuation technique must be choice-based [1–3]. The two choice-based techniques most commonly used to value preference-based measures are the cardinal methods of standard gamble (SG) and time trade-off (TTO) [4–6]. There are concerns about these cardinal methods because they are likely to be affected by factors other than a respondent’s preference for the state, such as risk aversion in the case of standard gamble or time preference and aversion to losses for TTO [7]. Furthermore, these tasks are cognitively complex and respondents might have some difficulty with them, particularly those in vulnerable groups such as the very elderly or children. For these reasons, there has been increasing interest in using ordinal tasks that require the respondent to rank one or more states [8–10] and in discrete choice experiments (DCE) involving pairwise comparisons [11–13].

J. Brazier · D. Rowen (✉) · A. Tsuchiya
Health Economics and Decision Science, School of Health and Related Research (ScHARR), University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA, UK
e-mail: d.rowen@sheffield.ac.uk

Y. Yang
Health Economics Research Group, Brunel University, London, UK

A. Tsuchiya
Department of Economics, University of Sheffield, Sheffield, UK

The ability to derive cardinal health state values from ordinal information comes from the assumption that a respondent's ranking of a set of states will be related to a latent utility variable. Individuals may give higher ranks to states with lower mean values than other states due to variability across individuals or random error. The proportion of occasions on which such reversals are made is related to the distance between mean values of the states in terms of the latent variable. There will be more agreement in rankings when the mean values for two states are further apart. The latent utility values are estimated using probabilistic choice models on ordinal data from multiple respondents.

A key problem in using ordinal methods has been how to anchor the values estimated by logistic models onto the full health-dead scale required for generating QALYs, anchoring full health at one and dead at zero. If the preference weights do not produce utility values on the full health-dead scale they cannot be used in economic evaluation using cost per QALY analysis. This paper uses existing anchoring techniques for rank and discrete choice data and presents an alternative anchoring technique for discrete choice data for two feasibility valuation studies; one for an asthma-specific measure and the other for an overactive bladder-specific measure. Preference weights obtained using rank and discrete choice data are compared to TTO results.

The paper begins by presenting an overview of the theory underlying ordinal methods. The methods and results of the valuation studies are presented, including a comparison of results using ranking, DCE and TTO on the same full health-dead scale. The implications of this study for further work are considered in the discussion.

Deriving cardinal values for health states from ordinal information

The idea of obtaining cardinal values from ordinal data first came from the work of Thurstone [14] who proposed the 'law of comparative judgement'. This was recognised [15] as offering a method for deriving cardinal preferences for health states from rank data and later implemented using the sleep dimension of the Nottingham Health Profile [8] and more recently the EQ-5D classification [16].

Thurstone's approach has been modified in a number of ways, including the application of a logistic function [17, 18] as a means of modelling the latent utility function from ordinal data. This uses two functions: one describes the probability of ranking one state over another given the utility of each state, and the second relates mean utility for each state to the severity levels for each dimension of the health state. Another important modification in this context is that in modelling a population level latent utility function from individual rank data, the error is characterised in

terms of the deviation of the individuals' preferences from population preferences. To use rank data, the assumption of independence from irrelevant alternatives (IIA) is required in order to explode the rank data into a series of pairwise choices, where ranking $A \succ B \succ C$ implies pairwise choices of $A \succ B$, $B \succ C$ and $A \succ C$. This assumes that the ordering of a pair of states does not depend on the other states being considered.

Recently, conditional logistic regression models were applied to the rank data collected as part of the UK valuation of the EQ-5D [9], SF-6D and HUI2 [10]. The rank model of health states alone does not produce utilities on the full health-dead scale necessary for use in generating QALYs, as it does not enable the anchoring of the values to 0 for dead. For this reason, the values generated by the logit model are transformed onto the full health-dead scale needed to generate QALYs. One method involves normalising the coefficients using the mean TTO value for the worst state defined by the classification system [9]. An alternative approach is to include the state 'dead' in the ranking exercise and normalise the regression coefficients so that 'dead' achieves a predicted value of zero [10].

DCE is a widely used tool in health economics for eliciting utility values, for example for different health care programmes, but has so far had limited use for eliciting health state values for preference-based measures of health used to derive QALYs. A limited number of studies have used DCE to value health states for their own sake [11–13, 19–21], but only one study has anchored their results onto the full health-dead scale required for generating QALYs. This study used a partial solution by normalising the DCE results using the estimated TTO value for the worst possible state [12]. The studies presented in this paper are the first attempt to undertake a normalisation of DCE results around dead without the use of cardinal values obtained from external sources. Here, we include the state 'dead' in the DCE and use this directly estimated parameter to rescale the regression coefficients. We compare the results to those obtained using the alternative approach of normalising using the estimated TTO value for worst state [12].

Methods

The health state classifications

Asthma-specific measure

The AQL-5D is a 5-dimension health state classification system [22] developed from the Asthma Quality of Life Questionnaire, AQLQ [23]. The dimensions of AQL-5D are: concern about asthma, shortness of breath, weather and pollution stimuli, sleep impact and activity limitations

(Table 1). The health state classification system has 5 dimensions each with 5 levels of severity, with level 1 denoting no problems and level 5 indicating extreme problems. By selecting one level for each dimension, it is possible to define 3,125 health states.

Overactive bladder-specific measure

The OAB-5D is a 5-dimension health state classification [24] developed from the overactive bladder instrument, OABq [25]. The dimensions of the OAB-5D are: urge, urine loss, sleep, coping and concern (Table 2). The health state classification system has the same structure as the AQL-5D, also defining a total of 3,125 health states.

The surveys

Two surveys were conducted for the each classification system. These surveys were identical in design in every way, apart from using different health state classifications

to define the health state descriptions. For each classification system, the surveys consisted of interviews containing ranking and TTO tasks and a follow-up postal survey using discrete choice tasks.

Interview

The interview surveys elicited values for a selection of states (AQL-5D/OAB-5D) from a representative sample of 300 members of the general public each. Adults who consented to participate were interviewed in their own home by an experienced interviewer trained by the authors of this paper. Respondents were asked to complete the health state classification questionnaire for themselves to help familiarise them with it. The first valuation task was to rank 7 intermediate states, full health (health state 11111), worst state defined by the health state classification ('pits' state 55555), and immediate death.

The next task was to value the 7 intermediate states and 'pits', with an upper anchor of full health using TTO. The

Table 1 Asthma quality of life classification (AQL-5D)

Concern

1. Feel concerned about having asthma none of the time
2. Feel concerned about having asthma a little or hardly any of the time
3. Feel concerned about having asthma some of the time
4. Feel concerned about having asthma most of the time
5. Feel concerned about having asthma all of the time

Short of breath

1. Feel short of breath as a result of asthma none of the time
2. Feel short of breath as a result of asthma a little or hardly any of the time
3. Feel short of breath as a result of asthma some of the time
4. Feel short of breath as a result of asthma most of the time
5. Feel short of breath as a result of asthma all of the time

Weather and pollution

1. Experience asthma symptoms as a result of air pollution none of the time
2. Experience asthma symptoms as a result of air pollution a little or hardly any of the time
3. Experience asthma symptoms as a result of air pollution some of the time
4. Experience asthma symptoms as a result of air pollution most of the time
5. Experience asthma symptoms as a result of air pollution all of the time

Sleep

1. Asthma interferes with getting a good night's sleep none of the time
2. Asthma interferes with getting a good night's sleep a little or hardly any of the time
3. Asthma interferes with getting a good night's sleep some of the time
4. Asthma interferes with getting a good night's sleep most of the time
5. Asthma interferes with getting a good night's sleep all of the time

Activities

1. Overall, not at all limited with all the activities done
 2. Overall, a little limitation with all the activities done
 3. Overall, moderate or some limitation with all the activities done
 4. Overall, extremely or very limited with all the activities done
 5. Overall, totally limited with all the activities done
-

Table 2 Overactive bladder quality of life classification system (OAB-5D)*Urge*

1. Not at all bothered by an uncomfortable urge to urinate
2. Bothered by an uncomfortable urge to urinate a little bit or somewhat
3. Bothered by an uncomfortable urge to urinate quite a bit
4. Bothered by an uncomfortable urge to urinate a great deal
5. Bothered by an uncomfortable urge to urinate a very great deal

Urine loss

1. Not at all bothered by urine loss associated with a strong desire to urinate
2. Bothered by urine loss associated with a strong desire to urinate a little bit or somewhat
3. Bothered by urine loss associated with a strong desire to urinate quite a bit
4. Bothered by urine loss associated with a strong desire to urinate a great deal
5. Bothered by urine loss associated with a strong desire to urinate a very great deal

Sleep

1. Bladder symptoms interfered with your ability to get a good night's rest none of the time
2. Bladder symptoms interfered with your ability to get a good night's rest a little of the time
3. Bladder symptoms interfered with your ability to get a good night's rest some of the time
4. Bladder symptoms interfered with your ability to get a good night's rest a good bit or most of the time
5. Bladder symptoms interfered with your ability to get a good nights rest all of the time

Coping

1. Bladder symptoms caused you to plan 'escape routes' to restrooms in public places none of the time
2. Bladder symptoms caused you to plan 'escape routes' to restrooms in public places a little of the time
3. Bladder symptoms caused you to plan 'escape routes' to restrooms in public places some of the time
4. Bladder symptoms caused you to plan 'escape routes' to restrooms in public places a good bit or most of the time
5. Bladder symptoms interfered with your ability to get a good night's rest all of the time

Concern

1. Bladder symptoms caused you embarrassment none of the time
2. Bladder symptoms caused you embarrassment a little of the time
3. Bladder symptoms caused you embarrassment some of the time
4. Bladder symptoms caused you embarrassment a good bit or most of the time
5. Bladder symptoms caused you embarrassment all of the time

survey used the TTO-prop method developed by the York Measurement and Valuation Health Group, which uses a 'time board' as a visual aid [26]. Respondents were then asked a series of socio-demographic questions. Finally, they were asked about their willingness to participate in a postal survey (described below).

The selection of health states for the interviews was determined by the specification of the model to be estimated. In this study, 98 health states and the worst state (to be repeated across the design) were selected out of the 3,125 possible health states described by the classification system. The selection was on the basis of a balanced design, which ensured that any dimension level (level λ of dimension δ) had an equal chance of being combined with all levels of the other dimensions. These 98 states were stratified into severity groups based on their total level score across the dimensions (simply the sum of the levels), and then randomly allocated into 14 blocks, so that each block has 7 health states. This procedure ensured that each respondent,

who was allocated one of the 14 blocks, received a set of states balanced in terms of severity and that each state is valued the same number of times except the worst possible state, the 'pits' state, which is valued by all respondents. Each state is valued by 20 respondents on average and this is comparable with other valuation studies, for example SF-6D states were valued by 15 respondents on average [5].

Postal surveys

A DCE questionnaire was mailed to interviewees who had consented to the postal survey approximately 4 weeks after the interviews (the 'warm' sample). Size of the warm sample depended on how many interviewed respondents were willing to participate in the postal survey. The same questionnaire was mailed out to a separate sample of the general public who had not been interviewed (the 'cold' sample'). The number of questionnaires mailed out was determined by targeted sample size and expected response

rate alongside funding constraints. Respondents were asked to complete the health state classification questionnaire for themselves to help familiarise them with it. Respondents were asked to indicate which state they preferred for an example pair of states and then for 8 pairs of states (see example question in the appendix). Finally, respondents were asked a series of socio-demographic questions. Reminders were sent to all non-responders approximately 4 weeks after the initial questionnaire was sent.

The large number of states defined by the classification systems of each measure mean it is infeasible to value all states. States were selected for the postal DCE using an application of a specially developed programme in the statistical package SAS [27], namely the D-efficiency approach. The programme obtains an optimal statistical design for DCE based on level balance, orthogonality, minimal overlap and utility balance. This reduces the number of pairwise comparisons to a manageable number. The programme produced 24 pairwise comparisons from the AQL-5D and OAB-5D, and these were randomly allocated to four versions of the questionnaire with 6 pairwise choices each. Two additional pairwise comparisons were included of two poor health states each compared to ‘immediate death’, and these were common across all versions of the questionnaire. No other states or pairwise comparisons were included in each version of the questionnaire. Only one pairwise comparison involves a logically consistent choice where one state has better health for every dimension.

Modelling health state values

Time trade-off

Time trade-off data was rescaled using the approach used in the UK EQ-5D value set [4] where worse than dead values are bounded at -1 . The data from the TTO valuation exercise was analysed using a one-way error components random effects model which takes account of variation both within and between respondents [5]. Estimation is via generalised least squares (GLS). The standard model is defined as:

$$U_{ij} = \beta(\mathbf{X}_{\delta\lambda}\boldsymbol{\beta}) + \varepsilon_{ij} \tag{1}$$

where $i = 1, 2, \dots, n$ represent individual health state values and $j = 1, 2, \dots, m$ represents respondents. The dependent variable U_{ij} is the disvalue (1-TTO value) for health state i valued by respondent j and $\mathbf{X}_{\delta\lambda}$ is a vector of dummy explanatory variables for each level λ of dimension $\hat{\delta}$ of the health state classification where level $\lambda = 1$ is the baseline for each dimension. ε_{ij} is the error term which is subdivided $\varepsilon_{ij} = u_j + e_{ij}$, where u_j is the individual random effect and e_{ij} is the usual random error term for the i th health state valuation of the j th individual. Details of other models run

on the TTO data are available elsewhere for both AQL-5D [28] and OAB-5D [29]. The value of the full health state equals 1 and health state values for all other states are estimated as 1 minus the coefficient for each of the appropriate level dummies for each dimension. The ‘gap’ between full health and the next best state is interpreted as the movement away from full health from having a problem on the appropriate dimension.

Ranking

The rank-ordered logit model was used to analyse the ranking data (a modelling approach also referred to as the conditional logit model [30]). It states that respondent j has a latent utility function for state i , U_{ij} and given the choice of two states i and k , the respondent will choose state j over state k if $U_{ij} > U_{ik}$.

The expected value of each unobserved utility was assumed to be a linear function of the categorical levels on the dimensions of the health state classification. Following the approach taken elsewhere [9, 10], the general model specification for each individual j 's cardinal utility function for state i is $U_{ij} = \mu_j + \varepsilon_{ij}$ where μ_j is the systematic component that is representative of the preferences of the population and ε_{ij} represents the specific preferences of the individual. The general model specification for analysis of the ranking data is:

$$U_{ij} = \beta(\mathbf{X}_{\delta\lambda}\boldsymbol{\beta} + \varphi D) + \varepsilon_{ij} \tag{2}$$

where U represents utility; $j = 1, 2, \dots, n$ represents respondents and $i = 1, 2, \dots, m$ represents health states. The functional form is assumed to be linear. The vector of dummies is as defined for Eq. 1, with the addition of a dummy variable for the state dead. For all health states other than dead $D = 0$. The rank-ordered logit model produces estimates on an interval scale, yet the origin and units of the interval scale are not on the full health-dead scale [9, 10]. In order to anchor onto the full health-dead scale, the coefficients relating to the levels of each dimension are normalised by dividing each level coefficient by the coefficient relating to dead; $\beta_{r\lambda\hat{\delta}} = \beta_{\lambda\hat{\delta}}/\Phi$ where $\beta_{r\lambda\hat{\delta}}$ is the rescaled coefficient for level λ of dimension $\hat{\delta}$, $\beta_{\lambda\hat{\delta}}$ is the coefficient for level λ of dimension $\hat{\delta}$ and Φ is the coefficient for dead [9, 10].

Discrete choice experiment

Two alternative approaches are used to obtain estimates onto the full health-dead scale. Method (1) models the DCE data using an existing approach in the literature [12]. This approach analyses the DCE data using a random effects probit model which takes account of the repeated measurement aspect of the data (whereby multiple responses are obtained from the same individual). This model excludes the data from the pairwise comparisons

involving ‘dead’. The value of the full health state is constrained to equal 1 and following the approach for all models estimated in this paper health state values for other states are estimated as 1 minus the coefficient for each of the appropriate level dummies for each dimension. The estimated coefficients are normalised onto the full health-dead scale using the estimated TTO value of the worst state. This means that the value of the worst state in the DCE model is anchored at the value of the worst state in the TTO model. Method (2) analyses all data from the DCE surveys including the pairwise comparisons involving dead using a random effects probit model. Again, an additive specification is used as specified by Eq. 2 where dead is included in the model. The coefficients are normalised in the same way as the rank data by dividing each level coefficient by the coefficient relating to dead. Models are also estimated separately for the ‘warm’ sample that was previously interviewed and the ‘cold’ sample that were not.

Comparison of models

All models are compared. There is no reason why rank or DCE models should produce the same results as the TTO model, although it could be thought that rank and DCE may produce similar results as the use of the rank-ordered

logit model means that the rank data is viewed as a series of pairwise comparisons.

Models can be compared in terms of the sign and ordering of their coefficients. The sign of the coefficients on the levels of each dimension are expected to be negative since they are all worse than the baseline (i.e. level 1). Furthermore, the levels in each dimension have a logical ordering, whereby more severe levels should have larger decrements. The number of inconsistencies between significant coefficients is compared between the models. For interest, we examine the relationship between model predictions and observed TTO values including the mean absolute difference, the root mean square of the difference, the proportions of differences greater than 0.05 and 0.1 and Pearson correlation coefficients. Finally, the pattern of the predictions is compared.

Results

Respondents

Three hundred and seven members of the public (response rate of 40%) in South Yorkshire (UK) were interviewed in the AQL-5D survey and 311 people interviewed in the OAB-5D survey (response rate of 26.7%). Table 3 shows

Table 3 Characteristics of respondents in valuation surveys

	AQL-5D n (%)	AQL-5D postal survey N (%)	OAB-5D N (%)	OAB-5D postal survey
Total	307	263	311	402
Age				
18–25	34 (11.1%)	9 (3.4%)	37 (11.9%)	14 (3.5%)
26–35	57 (18.6%)	35 (13.3%)	57 (18.3%)	47 (11.7%)
36–45	61 (19.9%)	45 (17.1%)	61 (19.6%)	71 (17.7%)
46–55	50 (16.3%)	56 (21.3%)	51 (16.4%)	81 (20.1%)
56–65	45 (14.7%)	64 (24.3%)	45 (14.5)	73 (18.2%)
> 66	60 (19.5%)	54 (20.5%)	60 (19.3%)	114 (28.4%)
Female	168 (54.7%)	148 (56.3%)	160 (51.4%)	236 (58.7%)
Married or living with partner	214 (69.8%)		217 (69.8%)	
Experienced serious illness				
In family	194 (63.4%)		176 (56.6%)	
In themselves	94 (30.6%)		94 (30.2%)	
Degree or equivalent	69 (22.5%)		85 (27.3%)	
Education after 17	140 (45.6%)		182 (58.5%)	
Renting property	64 (20.8%)		63 (20.2)	
Found valuation tasks in interview difficult				
Very difficult	24 (7.9%)		13 (4.2%)	
Quite difficult	82 (26.7)		80 (25.9%)	
Neither difficult nor easy	52 (16.9)		70 (22.7%)	
Self-reported EQ-5D scores				
Male, female	0.83, 0.84	0.81, 0.82	0.88, 0.88	0.87, 0.85

that the two samples were very similar in terms of their socio-demographic composition. Amongst the respondents to the AQL-5D survey, 53 (17.3%) had asthma and in the OAB-5D survey 27 (8.7%) reported experiencing symptoms of urge and 18 (5.8%) reported urine loss for at least some of the time. Overall self-reported health status using EQ-5D [4] was very close to the UK EQ-5D norms of 0.85 for females and 0.86 for males [31]. Two hundred and sixty three people responded to the AQL-5D postal survey and 402 people responded to the OAB-5D postal survey. Table 3 shows that the socio-demographic composition of the postal samples are similar to the interview samples, but the OAB-5D postal survey has a larger proportion of respondents over 65 years of age and a higher proportion of females. Overall, the AQL-5D samples have lower mean EQ-5D scores.

The data set

AQL-5D

There were 2,455 TTO health state valuations generated by the 307 respondents from the interviews and 3,041 rankings from the respondents at their interview. The average number of TTO valuations per intermediate health state was 22 (range from 19 to 22) and the worst possible state (AQL-5D state 55555) was valued by every respondent ($n = 307$). Mean TTO health state values ranged from 0.39 to 0.94 and generally have fairly large standard deviations (around 0.2–0.4). The distribution of the values was negatively skewed.

There were 168 DCE questionnaires returned out of the 308 who had been interviewed (55%) generating 1,336 observed pairwise comparisons. In total, 95 DCE questionnaires were returned in the cold survey (a 23% return rate) generating 741 pairwise comparisons.

OABq

There were 2,487 health state values generated by the 311 respondents and 3,040 rankings. Each intermediate health state was valued 22 times using TTO (range from 17 to 29) and the worst possible state (OAB-5D 55555) was valued 310 times using TTO (one missing value). Mean TTO health state values ranged from 0.56 for the worst possible state, to 0.91 for state 13,321, with an average standard deviation of 0.28.

The warm survey had 133 returned DCE questionnaires (response rate 44%) generating 1,050 pairwise comparisons. The cold survey resulted in 268 being returned (response rate 27%) generating 2,059 comparisons.

Modelling

AQL-5D

The TTO model and transformed rank and DCE models are presented in Table 4. The TTO model produced the expected negative coefficients for all statistically significant coefficients and the ordering of coefficients was consistent with the dimension levels of the AQL-5D. Three coefficients were positive but statistically insignificant. The rank model produced all negative coefficients and no inconsistencies for all significant coefficients. In comparison to the TTO and rank models, the DCE models have a higher number of positive coefficients and inconsistencies. The DCE results using method (1) that normalises coefficients using the estimated TTO value for the worst state has four positive coefficients, one of which is statistically significant, and one inconsistency between significant coefficients. The DCE results using method (2) for the pooled data (i.e. warm plus cold) produced three positive coefficients, one of which is statistically significant, and one inconsistency between significant coefficients. The warm DCE model produced five positive coefficients, none of which were statistically significant, and one inconsistency amongst statistically significant coefficients. The cold model had one positive coefficient that was not statistically significant and no inconsistencies between significant coefficients. The weather dimension seemed to cause most difficulty for the DCE models, with a suggestion that the levels of this dimension do not conform to the suggested ordering.

The size of the dimension level coefficients of the rank and TTO models are quite similar and follow an orderly pattern against the levels of the AQL-5D. The DCE model for the pooled data set reveals some marked differences. The most noticeable differences lie at the lower end of concern, short of breath, pollution and the upper ends of sleep and activity. Level 2 for the dimensions of concern, breath and pollution are all positive and in the wrong direction, quite markedly so for pollution. Sleep and activity have coefficients with the right sign, but they are much larger for levels 4 and 5.

The similarity of the TTO, rank and DCE models can be seen in the plot of predicted health state values against observed mean TTO values in Fig. 1. Mean absolute differences from observed TTO are 0.056 and 0.061 for the TTO and rank models, respectively, with mean differences of around zero. By contrast, the DCE predictions follow different paths depending on the normalisation method used. The DCE model that normalises coefficients using method (1) tended to have health state predicted values that were higher than observed TTO whereas the DCE model

Table 4 TTO and normalised rank and DCE model estimates^a for AQL-5D

Dimension level	TTO	Rank ^a	Discrete choice experiment			
			Method (1)		Method (2) ^a	
			Pooled data	Pooled data	Warm data	Cold data
Concern2	-0.028	-0.018	0.008	0.012	0.021	-0.006
Concern3	-0.044*	-0.043*	-0.015	-0.024	-0.006	-0.045
Concern4	-0.054*	-0.092*	-0.058*	-0.099*	-0.101*	-0.103*
Concern5	-0.081*	-0.127*	-0.096*	-0.139*	-0.123*	-0.164*
Breath2	0.000	-0.038*	0.025	0.025	0.044	-0.010
Breath3	-0.036*	-0.059*	-0.003	-0.008	0.004	-0.024
Breath4	-0.101*	-0.068*	-0.057*	-0.116*	-0.092*	-0.153*
Breath5	-0.116*	-0.106*	-0.093*	-0.138*	-0.128*	-0.147*
Pollution2	-0.019	-0.010	0.055*	0.084*	0.107*	0.046
Pollution3	-0.050*	-0.048*	0.010	-0.002	0.004	-0.006
Pollution4	-0.058*	-0.055*	-0.023	-0.051*	-0.049	-0.056
Pollution5	-0.121*	-0.071*	-0.063*	-0.085*	-0.095*	-0.060
Sleep2	0.018	-0.003	-0.027	-0.022	-0.025	-0.017
Sleep3	0.010	-0.016	-0.047*	-0.072*	-0.076*	-0.080
Sleep4	-0.033*	-0.047*	-0.094*	-0.125*	-0.104*	-0.165*
Sleep5	-0.054*	-0.068*	-0.100*	-0.149*	-0.117*	-0.199*
Activity2	-0.039*	-0.064*	-0.032*	-0.056*	-0.064*	-0.051
Activity3	-0.059*	-0.081*	-0.074*	-0.113*	-0.115*	-0.113*
Activity4	-0.175*	-0.163*	-0.158*	-0.247*	-0.262*	-0.232*
Activity5	-0.197*	-0.194*	-0.217*	-0.335*	-0.365*	-0.297*
Dead dummy		-1.000*		-1.000*	-1.000*	-1.000*
Number of observations	2,456	3,041	1,559	2,077	1,336	741
Number of individuals	307	306	263	263	168	95
Inconsistencies ^b	0	0	1	1	1	0
No. predictions > 0.05 from observed TTO	19	24	24	34	33	39
No. predictions > 0.1 from observed TTO	9	9	11	24	21	32
MAD from TTO	0.056	0.061	0.075	0.093	0.089	0.119
RMSD from TTO	0.070	0.079	0.093	0.118	0.111	0.149
Mean Error	-0.025	0.001	-0.060	0.059	0.036	0.102
Correlation with TTO	0.816	0.715	0.792	0.797	0.790	0.772

Method (1) normalises coefficients onto the full health-dead scale using estimated TTO value of worst state. Method (2) normalises coefficients onto the full health-dead scale using the dead coefficient

MAD from TTO: mean absolute difference between estimates and observed TTO values. RMSD from TTO: Root mean squared deviation between estimates and observed TTO values. Correlation with TTO: Pearson correlation coefficient between estimates and observed TTO

* Statistically significant at 5% level

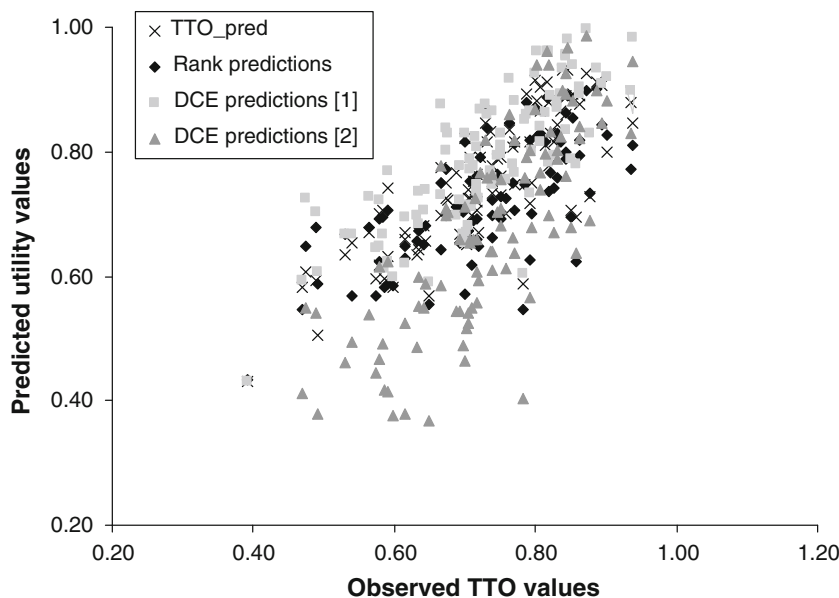
^a Normalised coefficient = estimated coefficient/dead dummy coefficient

^b Relating to statistically significant dimensions only

that rescaled coefficients using method (2) tended to have health state values lower than observed TTO values. The results from the DCE model using method (1) where coefficients are normalised using the estimated TTO value for worst state are more similar to the TTO model estimates, as expected due to the method of normalisation. Pearson correlation coefficients between observed TTO

values and predicted values for each model are consistently high and similar across models. Differences are observed between the mean values for the worst AQL-5D health state of 0.390 for observed TTO and predictions of 0.431 for TTO and DCE method (1), 0.434 for rank data and 0.154 for predictions from pooled DCE data normalised using method (2).

Fig. 1 Predictions of TTO, Rank and DCE models for AQL-5D in comparison to observed mean TTO



OAB-5D

The OAB-5D results are presented in Table 5. Overall, the models were broadly consistent with the ordinality of the OAB-5D. All the coefficients in the TTO model were negative and most significant. There were inconsistencies between significant coefficients in 3 cases, but their magnitudes were 0.02 or less. The ranking data produced negative coefficients and all but one were statistically significant with no inconsistencies between significant coefficients. The DCE results for method (1) had no inconsistencies between significant coefficients and has four positive coefficients, none of which were significant. All DCE models normalised using method (2) have five positive coefficients, one of which is statistically significant (coping level 2) and two inconsistencies amongst the significant coefficients.

The OAB-5D TTO model does not predict observed TTO as well as for the AQL-5D as indicated by mean absolute deviation (MAD) and mean error in Tables 4 and 5. Ranking predictions also do not agree with TTO as closely as for the AQL-5D survey and tended to have predicted health state values that are higher than observed TTO values. As for the AQL-5D survey, the DCE predictions normalised using method (2) have a larger scale range (0.249–1.00 compared to 0.623–1.0 for TTO and DCE method (1) and 0.436–1.0 for ranking). Again, the DCE models have different results depending on the method of normalisation. Again, the model estimated using method (2) to rescale coefficients tended to have predicted health state values lower than observed TTO, whereas the model estimated using method (1) tended to have predicted health state values higher than observed TTO, as shown in Fig. 2.

Pearson correlation coefficients between observed TTO values and predicted values for each model are high and similar across models, but are all lower than the equivalent values for the AQL-5D survey.

Discussion

This study has compared health state utility values on the full health-dead scale required to generate QALYs that have been derived using DCE, rank and TTO data. As would be expected, the TTO model best predicted TTO observed values, but then there is no reason to expect rank and DCE data to produce the same values. Perhaps more surprising is the way the rank model coefficients were actually very similar to TTO coefficients in the AQL-5D survey, but less so in the OAB-5D survey. In both surveys, the DCE model was the most different from the other methods, and the model normalising coefficients using the dead coefficient (method (2)) produced a larger range of values.

In modelling, rank data are essentially treated as a series of pairwise comparisons, and aside from the IIA assumptions, are otherwise the same. It is therefore interesting to find that they do not produce the same values. This may suggest that the rank and DCE tasks generate different data, which would have implications for the IIA assumption used in rank data. However, this may be simply due to differences in the study design and number and composition of health states valued. It may also reflect the fact that the ranking task preceded the TTO in the same interview, whereas the DCE data were collected via a postal survey on a different sample (although the warm DCE sample is composed of willing respondents from the interview).

Table 5 TTO and normalised rank and DCE model estimates for OAB-5D

Dimension level	TTO	Rank ^a	Discrete choice experiment			
			Method (1)		Method (2) ^a	
			Pooled data	Pooled data	Warm data	Cold data
Urge2	-0.033*	-0.065*	0.024*	0.048	0.072	0.034
Urge3	-0.026*	-0.086*	0.003	0.011	0.008	0.010
Urge4	-0.065*	-0.119*	-0.035*	-0.109*	-0.117*	-0.106*
Urge5	-0.083*	-0.178*	-0.063*	-0.169*	-0.154*	-0.175*
Urine2	-0.018	-0.028*	0.002	-0.023	-0.056	-0.012
Urine3	-0.049*	-0.039*	-0.012	-0.030	0.009	-0.050
Urine4	-0.030*	-0.060*	-0.043*	-0.134*	-0.061	-0.171*
Urine5	-0.041*	-0.093*	-0.046*	-0.091*	-0.098*	-0.090*
Sleep2	-0.027*	-0.027*	-0.004	0.000	-0.014	0.012
Sleep3	-0.019	-0.027*	-0.009	0.004	-0.040	0.032
Sleep4	-0.053*	-0.039*	-0.059*	-0.148*	-0.170*	-0.131*
Sleep5	-0.052*	-0.091*	-0.080*	-0.152*	-0.152*	-0.148*
Coping2	-0.004	-0.011	0.002	0.087*	0.117*	0.074*
Coping3	-0.018	-0.033*	-0.023*	-0.011	0.030	-0.028
Coping4	-0.021	-0.040*	-0.028*	-0.009	-0.008	-0.011
Coping5	-0.064*	-0.055*	-0.055*	-0.068*	-0.088*	-0.058
Concern2	-0.031*	-0.036*	-0.018*	-0.028	-0.029	-0.027
Concern3	-0.046*	-0.059*	-0.051*	-0.108*	-0.096*	-0.112*
Concern4	-0.085*	-0.095*	-0.095*	-0.235*	-0.244*	-0.231*
Concern5	-0.137*	-0.147*	-0.133*	-0.271*	-0.307*	-0.248*
Dead dummy		-1.000*		-1.000*	-1.000*	-1.000*
Number of observations	2,485	3,040	2,347	3,117	1,050	2,059
Number of individuals	311	304	402	402	133	268
Inconsistencies ^b	3	0	1	2	2	2
No. predictions > 0.05 from observed TTO	28	38	33	37	37	33
No. predictions > 0.1 from observed TTO	5	18	14	29	29	31
MAD from TTO	0.061	0.068	0.086	0.112	0.120	0.112
RMSD from TTO	0.073	0.086	0.100	0.142	0.152	0.141
Mean error	-0.043	0.042	-0.078	0.064	0.057	0.064
Correlation with TTO	0.698	0.619	0.670	0.649	0.614	0.640

Method (1) normalises coefficients onto the full health-dead scale using estimated TTO value of worst state. Method (2) normalises coefficients onto the full health-dead scale using the dead coefficient

MAD from TTO: mean absolute difference between estimates and observed TTO values. RMSD from TTO: Root mean squared deviation between estimates and observed TTO values. Correlation with TTO: Pearson correlation coefficient between estimates and observed TTO

* Statistically significant at 5% level

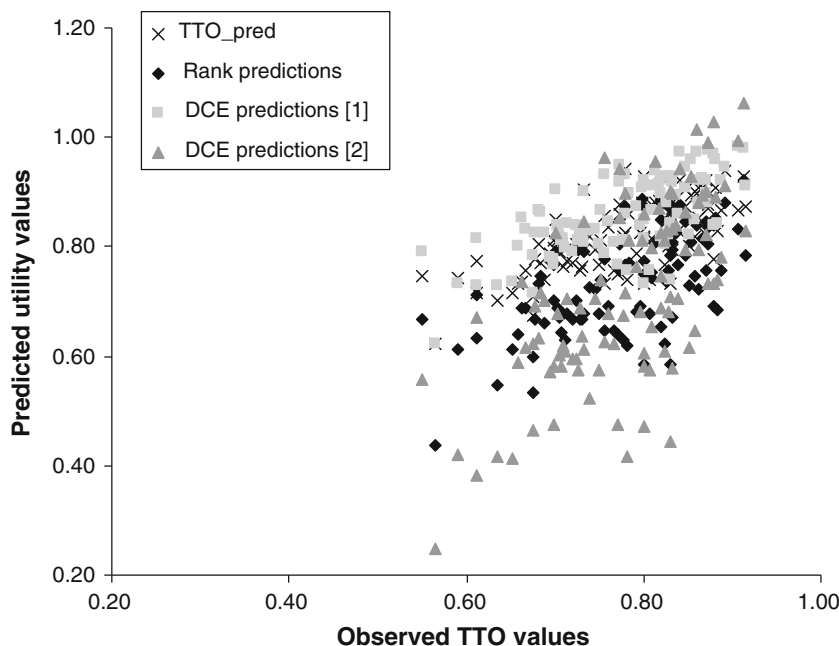
^a Normalised coefficient = estimated coefficient/dead dummy coefficient

^b Relating to statistically significant dimensions only

The DCE models based on the warm and cold samples seem to have similar coefficients and so were pooled to focus on the main comparisons with TTO and rank results and the existing approach (method (1)) used to anchor values onto the full health to dead scale [12]. Yet, the pooled data should be treated with some caution as further analysis did find some difference between the samples. However, the

sample sizes are small for the ‘cold’ and ‘warm’ samples, particularly for the cold AQL-5D sample. These results suggest the cold sample gave slightly lower values than the sample that had previously been interviewed, though this difference is not sufficiently large to alter the main findings comparing the different valuation methods. The similarity of warm and cold results suggests that it may be possible to

Fig. 2 Predictions of TTO, Rank and DCE models for OAB-5D in comparison to observed mean TTO



obtain DCE data to value health states without prior interview. This would be considerably cheaper, but postal surveys are usually associated with lower response rates and this was true for the AQL-5D survey. For researchers seeking to use DCE without other methods, it may still be preferable to approach respondents directly in their own home to ensure a more representative sample.

The pooled DCE models using different methods to rescale onto the full health-dead scale produce noticeably different coefficients and different ranges of predicted values. As expected the model normalising coefficients using the estimated TTO value of worst state (method (1)) is more similar to observed TTO values and the TTO model. Overall, the results suggest that DCE and TTO produce different results, and the use of TTO data to rescale DCE coefficients rather than using data collected using a DCE alone produces different results. This should be recognised in the future design of DCE surveys to obtain health state values.

The method used here to rescale worse-than-dead TTO values has been raised as a concern in the literature (see for example [32]) as negative values are bounded at -1 and may be interpreted as being measured on a different scale to TTO values that are better-than-dead. This may be considered as a limitation to the TTO model and to the model normalising DCE coefficients using the estimated TTO value of worst state (method (1)). However, this concern is likely to be of lower importance for these measures where only a small proportion of TTO responses are worse-than-dead, as 4% of TTO observations are worse-than-dead for the AQL-5D and only 2% of TTO observations are worse-than-dead for the OAB-5D.

The DCEs were feasibility studies added to valuation studies designed to provide TTO valuations of the AQL-5D and OAB-5D that are recommended by agencies such as NICE [2]. Using a postal method for DCE, for example, may have compromised the quality of the data and it certainly resulted in a lower response rate. The state selection used here did not ensure that implausible states were not chosen, but selected health states were checked to ensure they were plausible as implausible states may lead to an increase in the random variability in responses. Perhaps more importantly, the recommended approach for state selection and design for DCE experiments continually evolves [33], and our study may have benefited from recent improvements in DCE design.

There are concerns with the types of models estimated here since they make restrictive distributional assumptions about the coefficients. Of particular concern is that some orderings are logically determined. For example, suppose there is a health state pair: j and k , and $\mu_j - \mu_k = X$, say 0.2, on the latent variable scale standardised to 1 for full health and 0 for dead. The current approach to modelling ordinal data assumes that any two states that are apart from each other by X will have the same proportion of respondent's incorrectly ranking j over k . However, it is reasonable to assume that the probability of error will not only be a function of how apart the two states are, but also whether or not the two states have a logically determined ordering. Suppose there are two sets of health state pairs that are apart by X , where pair 1 has no logically determined ordering (e.g. 11122 and 33111) whereas pair 2 has a logically determined ordering (e.g. 11122 and 11133). It is reasonable to expect that the proportion of responses that

rank j over k will be different across pair 1 and pair 2. This becomes particularly problematic when one of the states is full health or the worst state. This means that the structure of the error term in Eq. 2 needs to be more sophisticated than it currently is. There is also a concern that the estimated parameters in the DCE model are confounded with an unknown scaling factor which is inversely related to the variance of the error term [34, 35]. However, this should not lead to biased coefficients. There are now more advanced econometric modelling techniques known as mixed logit models [36] that should be explored in future research. This would also overcome the IIA assumption underlying the way rank data are being analysed.

This paper presents a new way of anchoring health state values derived from discrete choice data on the full health-dead scale required for QALY estimation. Dead is included as a state in the pairwise choices and subsequently used to anchor the values generated by the logistic models. Another way to achieve this anchoring would be to include survival as a separate attribute [37]. However, this requires a far larger and more complex design, since survival has a multiplicative relationship to health related quality of life in the QALY model. This has been achieved through an online panel sample for EQ-5D [37]. One disadvantage with including dead as a state and using this to normalise coefficients arises from the fact that many respondents may not regard any state defined by the classification as worse than being dead and so effectively not be willing to trade [38]. This is likely to be more of a problem for milder descriptive systems. For these studies, a sufficient proportion of respondents were willing to make a trade, so that at the aggregate level it has been possible to estimate a societal value for the state of being dead compared to the health states defined by the health state classification.

This study proposes a new method for generating health state utility values on the QALY scale using discrete choice data. The results of feasibility studies on two condition-specific measures suggest that the relationship between health state utility values derived using TTO, rank and DCE data differs across different health state classification systems and potentially different medical conditions. Whilst ordinal methods may offer a promising alternative to conventional cardinal methods of SG and TTO, there is a large and important research agenda to address.

Acknowledgments We would like to thank the Centre for Research and Evaluation at Sheffield Hallam University for conducting the interviews. We are grateful to members of HESG for their comments on an earlier version of the paper. The studies reported in this paper were funded by Novartis (ALQ-5D) and Pfizer Inc. (OAB-5D). The studies reported in this paper were funded by Novartis (AQL-5D) and Pfizer Inc. (OAB-5D).

Appendix: Example question from the discrete choice survey

Health state A	Health state B
Bothered by an uncomfortable urge to urinate <i>a little bit or somewhat</i>	Bothered by an uncomfortable urge to urinate <i>a very great deal</i>
<i>Not at all</i> bothered by urine loss associated with a strong desire to urinate	Bothered by urine loss associated with a strong desire to urinate <i>a great deal</i>
Bladder symptoms interfered with your ability to get a good night's rest <i>none of the time</i>	Bladder symptoms interfered with your ability to get a good night's rest <i>some of the time</i>
Bladder symptoms caused you to plan 'escape routes' to restrooms in public places <i>none of the time</i>	Bladder symptoms caused you to plan 'escape routes' to restrooms in public places <i>some of the time</i>
Bladder symptoms caused you embarrassment <i>some of the time</i>	Bladder symptoms caused you embarrassment <i>a good bit or most of the time</i>

Which health state do you think is better? (please tick one box only)

A	B
<input type="checkbox"/>	<input type="checkbox"/>

References

1. Gold, M.R., Siegel, J.E., Russell, L.B., Weinstein, M.C.: Cost-Effectiveness in Health and Medicine. Oxford University Press, Oxford (1996)
2. NICE (National Institute for Health and Clinical Excellence): Guide to the Methods of Technology Appraisal. NICE, London (2008)
3. Drummond, M.F., Sculpher, M., O'Brien, B., et al.: Methods for the Economic Evaluation of Health Care Programmes. Oxford Medical Publications, Oxford (2005)
4. Dolan, P.: Modelling valuation for Euroqol health states. *Med. Care* **35**, 351–363 (1997)
5. Brazier, J., Roberts, J., Deverill, M.: The estimation of a preference based single index measure for health from the SF-36. *J. Health. Econ.* **21**, 271–292 (2002)
6. Feeny, D., Furlong, W., Torrance, G., et al.: Multiattribute and single attribute utility functions for the Health Utilities Index Mark 3 system. *Med. Care* **40**, 113–128 (2002)
7. Bleichrodt, H.: A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Econ.* **11**, 447–456 (2002)
8. Kind, P.: A comparison of two models for scaling health indicators. *Int. J. Epidemiol.* **11**, 271–275 (1982)
9. Salomon, J.A.: Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. *Popul. Health Metrol.* **1**, 12 (2003)
10. McCabe, C., Brazier, J., Gilks, P., et al.: Using rank data to estimate health state utility models. *J. Health. Econ.* **25**, 418–431 (2006)

11. Burr, J.M., Kilonzo, M., Vale, L., et al.: Developing a preference-based glaucoma utility index using a discrete choice experiment. *Optom. Vis. Sci.* **84**, 797–808 (2007)
12. Ratcliffe, J., Brazier, J., Tsuchiya, A., et al.: Using DCE and ranking data to estimate cardinal values for health states for deriving a preference-based single index from the sexual quality of life questionnaire. *Health Econ.* **18**, 1261–1276 (2009)
13. Ryan, M., Netten, A., Skatun, D., et al.: Using discrete choice experiments to estimate a preference-based measure of outcome—an application to social care for older people. *J. Health Econ.* **25**, 927–944 (2006)
14. Thurstone, L.L.: A law of comparative judgement. *Psychol. Rev.* **34**, 273–286 (1927)
15. Fanshel, S., Bush, J.W.: A health status index and its application to health services outcomes. *Oper. Res.* **18**, 1021–1066 (1970)
16. Kind, P.: Applying paired comparisons models to EQ-5D valuations—deriving TTO utilities from ordinal preferences data. In: Kind, P., Brooks, R., Rabin, R. (eds.) *EQ-5D Concepts and Methods: A Developmental History*. Springer, The Netherlands (2005)
17. Luce, R.D.: *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York (1959)
18. McFadden, D.: Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (ed.) *Frontiers in Econometrics*. Academic Press, New York (1974)
19. Hakim, Z., Pathak, D.S.: Modelling the EuroQol data: a comparison of discrete choice conjoint and conditional preference modelling. *Health Econ.* **8**, 103–116 (1999)
20. Johnson, R., Banzhaf, M., Desvousges, W.: Willingness to pay for improved respiratory and cardiovascular health: a multiple-format, stated preference approach. *Health Econ.* **9**, 295–317 (2000)
21. Osman, L.M., McKenzie, L., Cairns, J., et al.: Patient weighting of importance of asthma symptoms. *Thorax* **56**, 138–142 (2001)
22. Young, T., Yang, Y., Brazier, J., Tsuchiya, A.: The use of Rasch analysis in reducing a large condition-specific instrument for preference valuation: the case of moving from AQLQ to AQL-5D. *Med. Decis. Mak.* **31**, 195–210 (2011)
23. Juniper, E.F., Guyatt, G.H., Ferrie, P.J., et al.: Measuring quality of life in asthma. *Am. Rev. Resp. Dis.* **147**, 832–838 (1993)
24. Young, T., Yang, Y., Brazier, J., et al.: The first stage of developing preference-based measures: constructing a health-state classification using Rasch analysis. *Qual. Life Res.* **18**, 253–265 (2009)
25. Coyne, K., Revicki, D., Hunt, T., et al.: Psychometric validation of an overactive bladder symptom and health related quality of life questionnaire: the OAB-q. *Qual. Life Res.* **11**, 563–574 (2002)
26. MVH Group: *The Measurement and Valuation of Health: Final Report on the Modelling of Valuation Tariffs*. Centre for Health Economics, University of York, York (1995)
27. Huber, J., Zwerina, K.: The importance of utility balance in efficient choice designs. *J. Market. Res.* **33**, 307–317 (1996)
28. Yang, Y., Tsuchiya, A., Brazier, J., Young, T.: Estimating a preference-based index for a 5-dimensional health state classification for asthma derived from the asthma quality of life questionnaire. *Med. Decis. Mak.* **31**, 281–291 (2011)
29. Yang, Y., Brazier, J.E., Tsuchiya, A., et al.: Estimating a preference-based index from the Over Active Bladder questionnaire. *Value Health* **12**, 159–166 (2009)
30. Brazier, J., Ratcliffe, J., Salomon, J.A., et al.: *The Measurement and Valuation of Health Benefits for Economic Evaluation*. Oxford University Press, Oxford (2007)
31. Kind, P., Harman, G., Macran, S.: *UK Population Norms for EQ-5D*. Centre for Health Economics Discussion Series, University of York, York (1999)
32. Patrick, D.L., Starks, H.E., Cain, K.C., Uhlmann, R.F., Pearlman, R.A.: Measuring preferences for health states worse than death. *Med. Decis. Mak.* **14**(1), 9–18 (1994)
33. Louviere, J.J.: What you don't know might hurt you: some unresolved issues in the design and analysis of discrete choice experiments. *Environ. Resource Econ.* **34**, 173–188 (2006)
34. Flynn, T., Louviere, J.J., Peters, T.J., Coast, J.: Best-worst scaling: what it can do for health care research and how to do it. *J. Health Econ.* **26**, 171–189 (2011)
35. Swait, J., Louviere, J.J.: The role of the scale parameter in the estimation and comparison of multinomial logit models. *J. Market. Res.* **30**, 305–314 (1993)
36. Train, K.: *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge (2003)
37. Bansback, N., Brazier, J., Tsuchiya, A., Anis, A.: Using a discrete choice experiment to estimate societal health state utility values. *Health Economics and Decision Science Discussion Paper 10/03*, University of Sheffield, Sheffield (2010)
38. Flynn, T.N., Louviere, J.J., Marley, A.A.J., Coast, J., Peters, T.J.: Rescaling quality of life values from discrete choice experiments for use as QALYs: a cautionary tale. *Popul. Health Met.* **6**, 6 (2008)