Population Health Metrics

Research

Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data
Joshua A Salomon*

Address: Harvard School of Public Health, Boston, MA, USA

Email: Joshua A Salomon* - jsalomon@hsph.harvard.edu

* Corresponding author

Abstract

Background: In survey studies on health-state valuations, ordinal ranking exercises often are used as precursors to other elicitation methods such as the time trade-off (TTO) or standard gamble, but the ranking data have not been used in deriving cardinal valuations. This study reconsiders the role of ordinal ranks in valuing health and introduces a new approach to estimate interval-scaled valuations based on aggregate ranking data.

Methods: Analyses were undertaken on data from a previously published general population survey study in the United Kingdom that included rankings and TTO values for hypothetical states described using the EQ-5D classification system. The EQ-5D includes five domains (mobility, self-care, usual activities, pain/discomfort and anxiety/depression) with three possible levels on each. Rank data were analysed using a random utility model, operationalized through conditional logit regression. In the statistical model, probabilities of observed rankings were related to the latent utilities of different health states, modeled as a linear function of EQ-5D domain scores, as in previously reported EQ-5D valuation functions. Predicted valuations based on the conditional logit model were compared to observed TTO values for the 42 states in the study and to predictions based on a model estimated directly from the TTO values. Models were evaluated using the intraclass correlation coefficient (ICC) between predictions and mean observations, and the root mean squared error of predictions at the individual level.

Results: Agreement between predicted valuations from the rank model and observed TTO values was very high, with an ICC of 0.97, only marginally lower than for predictions based on the model estimated directly from TTO values (ICC = 0.99). Individual-level errors were also comparable in the two models, with root mean squared errors of 0.503 and 0.496 for the rank-based and TTO-based predictions, respectively.

Conclusions: Modeling health-state valuations based on ordinal ranks can provide results that are similar to those obtained from more widely analyzed valuation techniques such as the TTO. The information content in aggregate ranking data is not currently exploited to full advantage. The possibility of estimating cardinal valuations from ordinal ranks could also simplify future data collection dramatically and facilitate wider empirical study of health-state valuations in diverse settings and population groups.

Background

In population health measures and economic evaluations of health interventions, one essential input is a set of weights that reflect the relative value of time spent in different health states. These health-state valuations constitute the critical link between information on mortality and information on non-fatal health outcomes in summary metrics such as disability-adjusted life years or quality-adjusted life years [1,2]. There has been rising interest in recent years in collecting data on health-state valuations from diverse general population samples, in order to construct meaningful health measures that are consistent with common notions of health [3], and to conform to recommendations that economic evaluations adopt a societal perspective when they are intended to inform resource allocation decisions [4]. A variety of different methods have been proposed for eliciting health-state valuations in community surveys, including the standard gamble, time trade-off, person trade-off and visual analog scale [5-7]. Amidst debates over the most appropriate technique – with arguments for and against different methods based on economic theory [8], psychometric performance [9] and normative considerations [10] – empirical results from multi-method studies have demonstrated differences in the values inferred from the different methods, but have failed to produce consensus on a single preferred method [7,9,11-14].

While ordinal rankings have been incorporated in several major studies [15-17], the ranking of states typically constitutes a "warm-up" exercise for other modes of eliciting preferences; data on rankings generally have not been considered a suitable basis for developing cardinal valuations of health states. In other applications, by contrast, ordinal ranks and other discrete choice data have been used more widely in the derivation of interval-scaled values. Examples may be found in areas as diverse as consumer marketing [18], political science [19], transportation research [20] and environmental economics [21].

The conceptual basis for inferring cardinal values from ordinal responses has its origins in the pioneering work of Thurstone [22]. Assuming that observed rankings for a set of items are related to latent cardinal values that are distributed around the mean levels for each item, a person may give a higher rank to an item with a lower mean value due to individual variability or random error. The frequency of these rank inversions is related to the proximity of the mean values for different items on the latent scale. Mean values that are far apart, in other words, will produce greater agreement in rankings than mean values that are close together. This fundamental insight underlies a variety of related strategies for data collection and analysis, for example conjoint analysis [23] and binary choice

methods that have been used to estimate willingness to pay and standard gamble values from interval-censored data [24,25]. Methods for estimating cardinal values from categorical responses also rely on similar analytical models and have been used recently to derive health-state valuations [26].

This paper proposes a reconsideration of the use of ordinal rankings in the valuation of health states, presents a first application of a modeling strategy for health-state rankings based on the conditional logit model, and suggests avenues for further development of this approach. The objectives of this study were (1) to demonstrate how estimation of cardinal valuations may be undertaken using aggregate data on ordinal rankings and a standard set of statistical tools; and (2) to compare the predictive validity of a valuation model estimated from ordinal ranks with that of a widely-cited prior model estimated from time trade-off values.

Methods

Data

Data were collected in a general population survey in the United Kingdom reported previously [15], including 3,395 respondents interviewed in their homes using a standardized protocol [27]. These data are available to the public through the Data Archive [28]. The design and implementation of the survey have been described in detail elsewhere, and a number of different analyses of the data have been undertaken [29-33].

Health states in the survey were described using the EQ-5D descriptive system [34], which consists of one item for each of five dimensions (mobility, self-care, usual activities, pain/discomfort and anxiety/depression), with three possible levels on each dimension (Table 1). Conventional shorthand refers to a particular state by a 5-digit profile of the domain levels as ordered in Table 1. For example, 12321 would signify no problems walking about; some problems washing or dressing self; unable to perform usual activities, some pain or discomfort; and not anxious or depressed [34].

Respondents first described their own health using this system. They were then asked to rank order, from best to worst, 13 different hypothetical states described by EQ-5D profiles, plus outcomes labeled as "immediate death" and "unconscious," with the aid of index cards. For each respondent, the 13 hypothetical states were drawn from a set of 42 states included in the study. The rank exercise was followed by ratings of the same states using a visual analog scale. The final valuation task involved a series of time trade-off (TTO) questions for the 13 EQ-5D states, with respondents first indicating whether or not a given state was preferred to death, and then answering a

Table 1: Components of the EQ-5D descriptive system

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

sequence of hypothetical choices consisting of varying lengths of survivorship in different health states. Retest interviews were conducted with a sub-sample of 221 respondents approximately 10 weeks after the first interviews [15].

Model
Analysis of data on the ordinal rankings of health states was based on the random utility model attributed to Luce [35] and McFadden [36]. The specification requires two functions: firstly, a statistical model that describes the probability of ranking a particular health state higher than another given the (unobserved) cardinal utility associated with each health state; and secondly, a valuation function that relates the utility for a given health state to a set of explanatory variables, in this case the levels on the five dimensions of the EQ-5D instrument.

Note that the use of the term *utility* here does not imply a direct correspondence to the notion of expected utility derived under the von Neumann-Morgenstern axioms. Neither does the model demand that rankings be associated with utility in terms of preference satisfaction or consumer behavior. While the original model was formulated in reference to utility-maximizing economic agents, generalization to other applications allows interpretation of the latent construct that underlies observed choices to be determined by the content of the survey items, rather than the theoretical germs of the model. For example, if respondents were asked explicitly to rank order hypothetical health states in terms of perceived levels of "healthiness" rather than their own preferences, what is labeled as utility in the model would be more aptly described as a cardinal scale of *health*.

Statistical model
The random utility model was operationalized using the conditional logit regression model, which has also been referred to variously as the rank-ordered logit [19] or exploded logit model [37]. The following description of the model is adapted from previous applications in marketing research [37] and sociology [38].

Each respondent is observed to rank *J* states, with Y_{ij} denoting the rank given to state *j* by respondent *i* (and following the convention that 1 is the "highest" ranking). To simplify the notation, we describe the model as if all respondents valued the same set of *J* states, without loss of generality. It is assumed that respondent *i* has a latent utility value for state *j*, U_{ij}, that includes a systematic component and an error term:

$$U_{ij} = \mu_j + \varepsilon_{ij} \quad (1)$$

In the application described in this paper, only the attributes of the health state determine the systematic component of the latent utility, so μ is indexed only by *j*; in other words, a given health state has the same expected latent utility value across all respondents. A more general specification of the model would allow for systematic variation in latent utility values that depends on attributes of the respondent as well.

A respondent will rank state *j* higher than state *k* if $U_{ij} > U_{ik}$. Allowing for the stochastic element in the model, the probability of this ordering is given by:

$$\text{Prob}(U_{ij} > U_{ik}) = \text{Prob}(\varepsilon_{ij} - \varepsilon_{ik} < \mu_j - \mu_k) \quad (2)$$

If the error terms are assumed to be independent and identically distributed with an extreme value distribution, given by $\text{Prob}(\varepsilon_{ij} \leq t) = \exp\{-\exp(-t)\}$, then the odds of ranking j higher than k simplify to $\exp\{\mu_j - \mu_k\}$, and the likelihood for the complete ordering of a particular respondent may be written as

$$L_i = \prod_{j=1}^{J} \left[\frac{\exp\{\mu_j\}}{\sum_{k=1}^{J} \delta_{ijk} \exp\{\mu_k\}} \right] \quad (3)$$

where $\delta_{ijk} = 1$ if $Y_{ik} \geq Y_{ij}$, and 0 otherwise (cf. [38]).

The extreme value distribution is a convenient option for the joint distribution of the error terms because it offers a simple closed-form expression for the choice probabilities. Given two variables X and Y with extreme value distributions, the difference $X - Y$ has a logistic distribution, hence the logit regression model. While other alternatives are possible, options such as the multinomial probit would require evaluation of complex integrals [37,38].

The name *exploded logit* has been used to describe the model because an observed rank ordering of J alternatives may be regarded as an "explosion" into $J - 1$ independent observations, such that $U_{i1} > U_{i2} > ... U_{iJ}$ gives rise to $(U_{i1} > U_{ij},\ j = 2,...,J)$, $(U_{i2} > U_{ij},\ j = 3,...,J)$, ..., $(U_{i(J-1)} > U_{iJ})$ [37]. Thus, the rank data are treated as equivalent to a sequence of choices, in which the state with the best rank is chosen over all other alternatives, the state with the second rank is chosen over all except the first, and so on. This explosion is made possible by the assumption of independence from irrelevant alternatives (IIA), which states that the ordering of a given pair of items does not depend on the other alternatives available [35].

In the present context, μ_j in equation 3 may be understood as the average valuation of a particular health state, and we may elaborate the model to express μ as a function of the multiple domain levels in the descriptive system, i.e., to specify the form of an EQ-5D valuation function as detailed in the following section.

Valuation function
A range of different specifications are possible for the valuation function that relates the utility of a given health state to levels on different domains of health. While many important conceptual and methodological issues around the specification of valuation functions continue to be debated in the literature [33,39-41], these considerations are not the main focus of this paper, so the analysis reported here does not include a comprehensive examination of alternative functional forms. Because the primary

aim is to demonstrate the usefulness and feasibility of a new approach to modeling cardinal valuations based on ordinal ranking data, a model analogous to a widely-cited previous model estimated from the TTO values in the same dataset [30] is adopted as a starting point, to facilitate comparison.

In the model, the expected value for the latent utility of each health state is assumed to be a linear function of the categorical ratings on the five EQ-5D domains:

$$\mu_j = x_j' \boldsymbol{\theta} \quad (4)$$

with x_j a vector of indicator variables referring to domain levels (Table 2) and θ a vector of unknown parameters. The model used here is algebraically equivalent to the model reported by Dolan [30], although specified slightly differently. In the Dolan model, the first set of variables for the dimension levels are equal to 1 if the dimension takes level 2; 2 if the dimension takes level 3; and 0 otherwise, while the second set of variables are equal to 1 if the dimension takes level 3, and 0 otherwise. In the present study, the first set of variables are equal to 1 if the dimension takes level 2 or level 3, and 0 otherwise; the second set of variables follows the Dolan specification. Thus, in the Dolan model, the contribution of a level 3 rating on a particular dimension would be twice the first coefficient plus the second, while in the present model the level 3 contribution is the sum of the two coefficients. The modification simplifies subsequent rescaling by allowing the valuation of the state characterized by the worst levels on all dimensions (i.e. the 33333 state) to be computed as the sum of all of the coefficients.

Scaling
The conditional logit model produces estimated valuations on an interval scale, such that meaningful comparisons of differences are possible [42]. However, the origin and units of the scale are defined arbitrarily by the identifying assumptions in the model. In other words, the rank order of a set of health states will be the same under any positive affine transformation of the latent utilities, which implies the following more general specification of equation 1 (cf. [37]):

$$U_{ij} = \alpha(\mu_j + \varepsilon_{ij}) + \beta \quad (5)$$

Substituting from equation 4, the predicted utility for a given health state, conditional on the parameter values estimated in the model, would be $\alpha x_j' \hat{\boldsymbol{\theta}} + \beta$.

In the context of health-state valuations, there are certain conceptual constraints on the possible values for the parameters α and β, which lead to a limited number of

Table 2: Variable definitions in the conditional logit regression model

Variable	Definition
M2	1 if mobility is at level 2 or higher; 0 otherwise
S2	1 if self-care is at level 2 or higher; 0 otherwise
U2	1 if usual activities is at level 2 or higher; 0 otherwise
P2	1 if pain/discomfort is at level 2 or higher; 0 otherwise
A2	1 if anxiety/depression is at level 2 or higher; 0 otherwise
M3	1 if mobility is at level 3; 0 otherwise
S3	1 if self-care is at level 3; 0 otherwise
U3	1 if usual activities is at level 3; 0 otherwise
P3	1 if pain/discomfort is at level 3; 0 otherwise
A3	1 if anxiety/depression is at level 3; 0 otherwise
N3	1 if any domain is at level 3; 0 otherwise

Note: This specification is algebraically equivalent, but not identical, to that used by Dolan [30] in modeling time trade-off values. The slight modification from Dolan is adopted so that the modeled value for the EQ-5D state 33333 is simply equal to the sum of all of the estimated coefficients from the model, which allows for convenient rescaling (see text).

logical alternatives. As applied here, β represents the value assigned to a state characterized by the best possible levels on all of the health dimensions in EQ-5D (i.e. the 11111 state). Intuitively, β = 1 is a reasonable choice that implies that a person with no difficulties on any dimension will have an expected health state valuation of 1. The choice of β = 1 is consistent with the Dolan model [30]; although the latter specification includes an estimated intercept term, the intercept does not apply to the 11111 state. Dolan interprets the intercept as an indication that "any move away from full health [is] associated with a substantial loss of utility" and notes that "[the intercept] could represent a discontinuity in the model between level 1 and level 2 in the much the same way as the 'N3' term represented a discontinuity between level 2 and level 3." (p. 1104)

For the value of α, which defines a normalizing constant for the model coefficients, there are a somewhat larger number of possibilities. Three alternatives are considered:

Normalization to match the scale of observed TTO values in the data
The modeled value for the 33333 state, on the untransformed scale, equals the sum of all of the estimated coefficients in the regression model, denoted by $\sum \hat{\theta}$. To transform the scale such that the 33333 state has a value equal to the mean observed TTO value for this state, denoted by $\overline{TTO}_{33333}$, we substitute β = 1 and $\mu_j = \sum \hat{\theta}$ in equation 5,

$$\alpha \sum \hat{\theta} + 1 = \overline{TTO}_{33333} \tag{6}$$

and solve for α:

$$\alpha = \frac{\overline{TTO}_{33333} - 1}{\sum \hat{\theta}} \tag{7}$$

Normalization to produce a utility of 0 for the 33333 state
Similarly, the value of α may be chosen to define a scale with the 33333 state having a utility of 0:

$$\alpha \sum \hat{\theta} + 1 = 0 \tag{8}$$

$$\alpha = -\frac{1}{\sum \hat{\theta}} \tag{9}$$

Normalization to produce a utility of 0 for death
For the third rescaling option, the observed rankings of death are added to the dataset, and an extended model is estimated including all of the variables in Table 2 plus an additional indicator variable, λ, which takes the value 1 for all observations pertaining to death, and 0 otherwise. Thus, λ represents the modeled utility for death on the same untransformed scale as the coefficients used to model the utilities of the EQ-5D states. As before, the value of α is determined by substituting in equation 5 and then solving:

$$\alpha \hat{\lambda} + 1 = 0 \tag{10}$$

$$\alpha = -\frac{1}{\hat{\lambda}} \tag{11}$$

A critical issue relating to rescaling is the interpretation of states worse than death. Various normative arguments may be made regarding the possibility of states worse than death; these arguments depend in some part on the definition of the quantity of interest in a particular study. As a

model of individual preferences, the possibility that death may be preferred to certain states is plausible, while a consideration of levels of health may be less accommodating to the notion of states worse than death – it is hard to imagine what it means to be "less healthy" than one who is dead. The choice over rescaling options ultimately depends on these normative arguments in addition to empirical considerations. As this paper is primarily an empirical investigation of a new method for modeling health-state valuations, however, the main comparison of the different scaling alternatives will focus on goodness-of-fit to observed data in the study.

Reversibility

The extreme value distribution is right-skewed, and as a result the exploded logit model does not give perfectly symmetric results when rank orderings are inverted. In other words, if states are ranked from best to worst in one analysis, an alternative analysis of rankings from worst to best would *not* produce coefficients that are identical but for opposite signs [38,43,44]. While this property may be unappealing intuitively, in practice the difference is usually minimal [43]. In order to consider whether the lack of reversibility produces substantively important differences in this case, the analysis has been run with inverted rank orderings as well for purposes of comparison.

Model evaluation and comparison

The principal objective of this paper, to assess the validity of a new approach to modeling health-state valuations, was pursued through comparison of predictions from the rank-ordered regression model to observed TTO values in the same dataset, and to predictions based on a previously reported model of directly-elicited TTO values [30]. For the rank model, predictions were computed for the 42 states included in the study as $\alpha x_j' \hat{\theta} + 1$ under the three alternative choices for the value of α. Predictive validity was assessed in terms of the intraclass correlation coefficient (ICC) between modeled values and mean observed TTO values for the 42 states, and the root mean squared errors (RMSE) of the predictions at the individual level.

Results
Descriptive analysis

Characteristics of the study population have been reported elsewhere [15,32]. It is useful, however, to begin with some brief descriptive analyses of the data. First, an examination of the test-retest reliability of the ranking and TTO questions offers insight into the degree of measurement error inherent in the two methods. For the 211 respondents who completed retest interviews, the ICCs (between test and retest, calculated for each individual) were higher on average and less variable for rankings than for TTO values (Table 3). Comparison of the ICCs
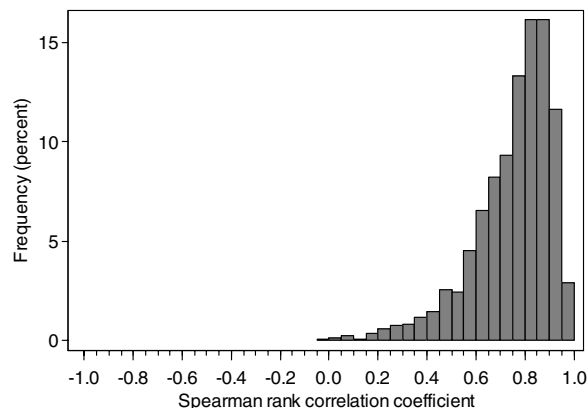
Figure 1
Spearman rank correlation coefficients for ordinal rankings and time trade-off values.

between ranks and TTO values may be complicated somewhat by the fact that ranking allows a smaller number of possible values than TTO, which might artificially minimize differences between test and retest responses. In order to account for this possibility, ICCs were also computed on the ordinal ranks implied by TTO values to equalize the advantage conferred by having few discrete values. The reliability results for the TTO-implied ranks remained lower than those for the direct rankings, which confirms that there is considerably more measurement error inherent in the TTO, such that even at the ordinal level TTO values are less reproducible than rankings elicited directly.

In light of the different measurement characteristics, it is worth investigating the overall level of agreement between the ordering of states in the ranking exercise and the TTO. Figure 1 shows the distribution of Spearman rank correlation coefficients between the rankings and TTO values in the full sample. The mean correlation coefficient was 0.78, and the median was 0.82. Given the findings on test-retest reliability, it is likely that the difference between direct rankings and implied TTO rankings is due in large part to measurement error, with the notable exception of the rank assigned to death.

The outcome of death was atypical in that 82% of respondents ranked death higher on the TTO than in the ordinal ranking exercise, with an average difference of 3.3 ranks between the TTO and direct rankings of death in the full sample. Excluding death, there was no other state with a mean absolute difference greater than 1.2 between the

Table 3: Test-retest reliability of rankings and time trade-off values

| | Intraclass correlation coefficient* | | |
Response type	Median	Mean	Standard deviation
Ordinal ranks	0.93	0.90	0.11
Time trade-off	0.79	0.74	0.19
Implied ranks from time trade-off	0.84	0.78	0.17

*Intraclass correlation coefficients between test and retest responses were computed at the individual level for 211 respondents.
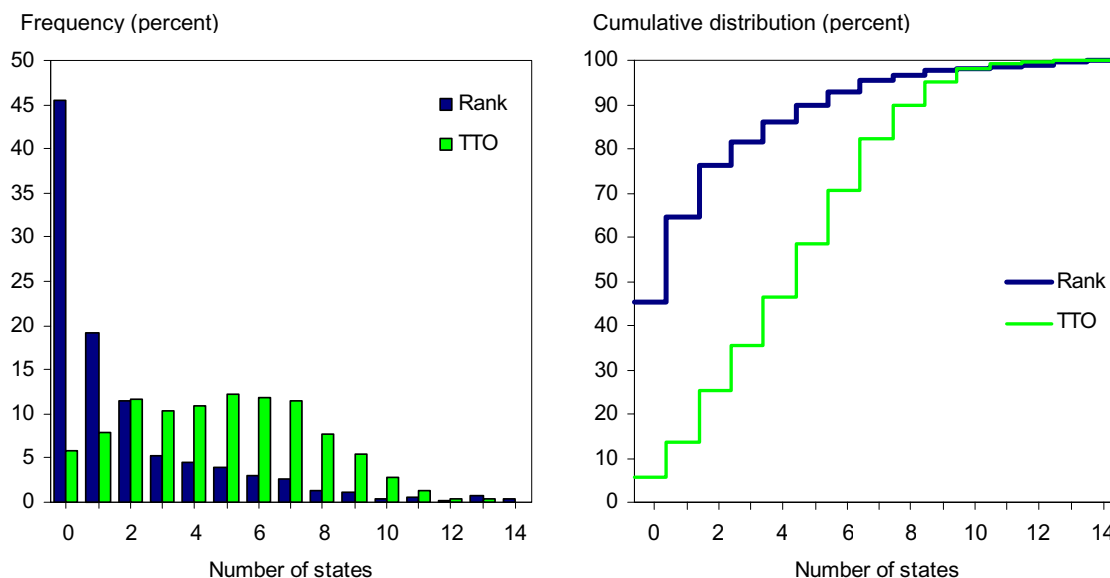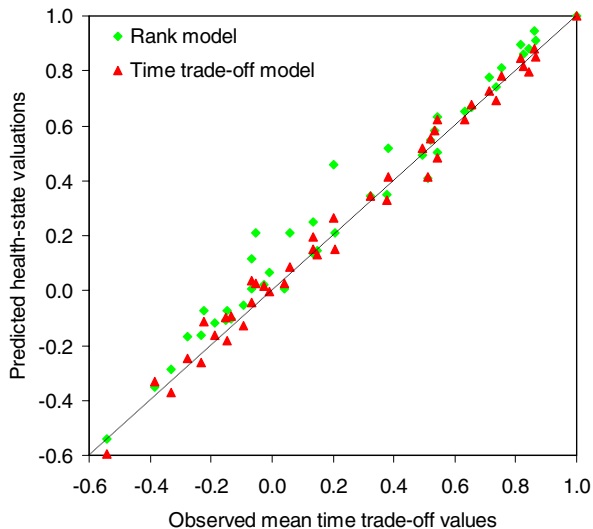

Figure 2
Number of states worse than death in rankings and time trade-off (TTO).

two sets of rankings. Considering the averages for each state across all respondents, only one state would be regarded as worse than death on rankings, compared to 16 on the TTO. At the individual level, the mean and median numbers of states rated worse than death in direct rankings were 1.8 and 1, respectively, while the mean and median numbers of TTO values worse than death were 4.8 and 5 (Figure 2). The significance of the different rank positions of death in the two methods will be revisited below in considering different scaling alternatives.

Results from the conditional logit regression model
Table 4 shows estimated coefficients from the conditional logit regression model of the rank data, as well as rescaled coefficients under the three alternatives described above: (Option 1) normalization to match the empirical TTO value of the 33333 state; (Option 2) normalization to set the utility of the 33333 state to 0; (Option 3) normaliza-

tion to set the utility of death to 0, based on the estimated coefficient for death in an extended model. The three alternative sets of predictions were each strongly correlated with the observed mean TTO values: Pearson's r was 0.985 for both option 1 and option 2 – by definition, linear transformations of one another; and 0.984 for option 3 – which deviates slightly from linearity with the other two because a separate model was estimated including the indicator variable for death. Using the ICC, on the other hand, which responds to both strength of association and mean differences, the rescaling by the lowest observed TTO emerged as the best-fitting alternative, with an ICC of 0.974 compared to 0.572 or 0.595 for options 2 and 3, respectively. Most notably, the fit of this rank model was only marginally lower than the fit for predictions based on the directly estimated TTO tariff function reported previously by Dolan [30], which gave an ICC of 0.993 (Figure 3).

Figure 3
Predicted and observed health-state valuations.
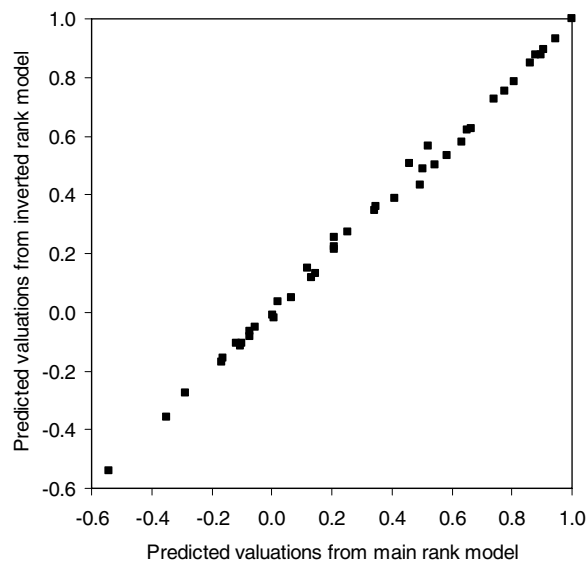
Figure 4
Comparison of predictions in main rank model and inverted model.

The difference between the fit of predicted values scaled using the lowest observed TTO versus those scaled by setting the value for death highlights the importance of the different findings regarding states worse than death in the ordinal ranking compared with the TTO responses. As noted above, death was unique in the degree to which its relative position shifted in the TTO exercise compared to the initial ranking. We may speculate that the difference is attributable in some way to the script that was used to elicit a categorization of states as better or worse than death at the outset of the TTO exercise, but important questions regarding these differences remain unresolved. For the purpose of this paper, however, the key finding is that predicted health-state valuations derived from a model of ordinal ranking data can provide a close match to observed differences between cardinal values for different states – i.e., the model may be used to generate robust predictions *on an interval scale*, with predictive validity rivalling that of a model estimated directly from TTO values.

Table 5 presents the comparison between modeled and observed TTO values by EQ-5D state, including the predictions from both the rank regression model normalized to the TTO scale and the previous TTO-based model. Among the four states with the largest discrepancies between predicted values from the rank model and mean observed TTO values, all included level 3 on the dimension of pain, and all were among the states with the largest

differences in rank positions between the direct ordering exercise and the TTO. Across all states, the mean absolute difference between predicted values and observed TTO values was 0.067 for the rank model, compared to 0.040 for the TTO model. At the individual level the errors were also comparable, with root mean squared errors of 0.502 and 0.496 for the rank and TTO models, respectively.

To consider the implications of asymmetry in the extreme value distribution, an alternative model was estimated based on inverting the ranks, such that greater rank numbers would correspond to higher utilities. Figure 4 shows a comparison of the predictions from the main model and the inverted model. The agreement between the two models was high, with an ICC of 0.998. Across the 42 states, the mean absolute difference between the predictions in the two models was 0.020, with a maximum of 0.062. Predictive validity of the inverted rank model compared to observed TTO values was almost identical to that of the main model (ICC = 0.967 in comparison to mean TTO observations), as was the average error at the individual level (RMSE = 0.503).

Discussion
This paper introduces a new approach to modeling health-state valuations based on ordinal rankings that produces robust predictions of observed valuations elic-

Table 4: Regression results from the conditional logit model for health-state rankings.

Variable	Coefficient	Std. Err.	p-value	Rescaled coefficients		
				$33333 = \overline{TTO}_{33333}$	$33333 = 0$	Death = 0
M2	-0.580	0.017	<0.001	-0.092	-0.059	-0.063
S2	-0.875	0.018	<0.001	-0.138	-0.090	-0.097
U2	-0.334	0.019	<0.001	-0.053	-0.034	-0.033
P2	-0.757	0.017	<0.001	-0.120	-0.078	-0.083
A2	-0.660	0.017	<0.001	-0.104	-0.068	-0.070
M3	-1.480	0.027	<0.001	-0.234	-0.152	-0.134
S3	-0.783	0.023	<0.001	-0.124	-0.080	-0.066
U3	-0.690	0.022	<0.001	-0.109	-0.071	-0.061
P3	-1.209	0.025	<0.001	-0.191	-0.124	-0.107
A3	-0.909	0.023	<0.001	-0.144	-0.093	-0.078
N3	-1.464	0.032	<0.001	-0.232	-0.150	-0.186

ited through the time trade-off technique. While ordinal rankings at the individual level do not indicate strength of preferences, the estimation of plausible valuations on an interval scale is nevertheless possible via models of aggregate-level data on rankings. In fact, the results in this study suggest that the information content of aggregate rank data is similar to that of data on widely recommended valuation methods such as the TTO. Although the degree of similarity is rather surprising, the fundamental intuition behind the extraction of cardinal values from aggregate rank data is straightforward: large cardinal differences are expected to produce greater agreement across respondents in the ordering of a particular pair of states than will small differences, and this principle extends easily to full rank sets.

It will be useful to confirm the results from this study in other surveys, and comparison to other widely-used methods such as the standard gamble may be instructive. A convenient starting point would be other datasets that have already been collected and analysed. Because ranking exercises have been included in several previous valuation studies, a number of comparisons similar to the one described in this paper might be made with minimal effort. The promising findings in this first application should also encourage the inclusion of ordinal ranking exercises in other planned surveys on health-state valuations, if they are not already incorporated in the protocols.

In the meantime, further methodological work will be useful in several areas. One important consideration is the assumption of independence from irrelevant alternatives that gives rise to the conditional logit formulation. The possibility that utilities are correlated across health states at the individual level should be considered through elaborations of the basic model described here. Other options for relaxing the IIA assumption are also worth exploring, for example allowing for larger random errors associated with later rankings in comparison to early ones [38]. The specification of the valuation function is a critical question that is not considered thoroughly in this study but warrants greater attention. One specific avenue of research that has stimulated rising interest in recent years is the question of potential variation in valuation functions within and between populations, which could be readily accommodated in the conditional logit model described here.

A potential limitation of the models for rank-ordered data that must be emphasized is the need to determine the scale of the unobserved utilities, as ranks are invariant under positive affine transformations of the underlying scale. It will be useful to consider the choice of scale anchors during the design phase of studies, as particular choices will have implications for the states that must be included in the ranking exercise. In the dataset used here, the scaling question was complicated by important empirical differences in the relative ranking of death in the time trade-off exercise compared to the direct ordinal ranks. A more detailed investigation of the determinants of this discrepancy is needed, as is further conceptual exploration of the notion of states worse than death. Nevertheless, the number of logical alternatives to define the scale of estimated valuations is limited, and both empirical investigation and normative reasoning may be brought to bear on a comparison of available options. While a rescaling in reference to the lowest observed time trade-off value provided the best-fitting predictions in the example described in this study, fixing the scale in reference to death may be an appealing option for other reasons. Issues regarding rescaling merit careful consideration in subsequent applications of this approach.

Table 5: Comparisons of observed and predicted valuations by state.

State	Observed*	Rank model		Time trade-off model†	
		Predicted	Difference	Predicted	Difference
21111	0.864	0.908	-0.045	0.850	0.014
11211	0.860	0.947	-0.088	0.883	-0.023
11121	0.841	0.880	-0.039	0.796	0.045
12111	0.823	0.862	-0.039	0.815	0.008
11112	0.818	0.896	-0.078	0.848	-0.030
12211	0.754	0.809	-0.055	0.779	-0.025
12121	0.734	0.742	-0.008	0.692	0.042
11122	0.712	0.776	-0.064	0.725	-0.013
22112	0.656	0.666	-0.010	0.675	-0.019
22121	0.634	0.650	-0.017	0.623	0.011
21222	0.543	0.631	-0.089	0.620	-0.077
11312	0.542	0.502	0.040	0.485	0.057
12222	0.534	0.585	-0.051	0.585	-0.051
22122	0.518	0.546	-0.027	0.552	-0.034
21312	0.511	0.411	0.101	0.416	0.095
22222	0.495	0.493	0.002	0.516	-0.021
11113	0.383	0.520	-0.137	0.414	-0.031
13212	0.377	0.349	0.028	0.329	0.048
13311	0.323	0.344	-0.021	0.342	-0.019
12223	0.206	0.209	-0.004	0.151	0.055
11131	0.202	0.458	-0.256	0.264	-0.062
21323	0.149	0.147	0.001	0.128	0.021
32211	0.135	0.252	-0.117	0.196	-0.061
23321	0.134	0.133	0.001	0.150	-0.016
21232	0.058	0.209	-0.151	0.088	-0.030
22323	0.042	0.009	0.033	0.024	0.018
22331	-0.009	0.066	-0.075	-0.003	-0.006
33212	-0.025	0.023	-0.049	0.015	-0.040
11133	-0.054	0.209	-0.263	0.028	-0.082
21133	-0.068	0.118	-0.186	-0.041	-0.027
23313	-0.069	0.005	-0.073	0.037	-0.106
23232	-0.094	-0.053	-0.040	-0.126	0.032
33321	-0.133	-0.101	-0.032	-0.095	-0.038
22233	-0.146	-0.073	-0.072	-0.181	0.035
32313	-0.153	-0.106	-0.047	-0.098	-0.055
32223	-0.186	-0.116	-0.070	-0.163	-0.023
13332	-0.226	-0.071	-0.155	-0.115	-0.111
32232	-0.231	-0.164	-0.067	-0.261	0.030
32331	-0.276	-0.168	-0.107	-0.248	-0.028
33232	-0.331	-0.288	-0.044	-0.371	0.040
33323	-0.383	-0.349	-0.034	-0.331	-0.052
33333	-0.540	-0.540	0.000	-0.594	0.054

*Mean observed time trade-off value †Time trade-off model based on Dolan [30]

Conclusions

The empirical basis for understanding health-state valuations in the general community has been limited to date, particularly in developing countries. One major constraint to expansion of the evidence base on valuations has been the complexity of the recommended tools for data collection, which in most cases demand abstract and cognitively challenging thought experiments on the part of the survey respondent. In contrast to techniques such as the standard gamble and time trade-off, on the other hand, ordinal ranking exercises represent a relatively simple means of data collection that – as shown in the present study – provide results that are highly reliable in test-retest situations. Most significantly, the findings in this study suggest that the information content in ordinal rankings has not been exploited to full advantage and point to encouraging new directions in data collection and analysis on health-state valuations. If these findings

are confirmed in other datasets, the possibility of estimating cardinal valuations from ordinal ranks might simplify future research on health-state valuations dramatically and facilitate wider empirical study of valuations in diverse settings and population groups.

List of abbreviations

TTO time trade-off

ICC intraclass correlation coefficient

RMSE root mean squared error

Competing interests

None declared.

Acknowledgments

References

1. Nord E: **Methods for quality adjustment of life years.** *Soc Sci Med* 1992, **34**:559-569.
2. Murray CJL: **Rethinking DALYs.** In: *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020.* Edited by: *Murray CJL and Lopez AD.* Boston, Harvard School of Public Health; 1996:1-98.
3. Salomon JA, Mathers CD, Chatterji S, Sadana R, Üstün TB, Murray CJL: **Quantifying individual levels of health: definitions, concepts and measurement issues.** In: *Health systems performance assessment: debates, methods and empiricism.* Edited by: *Murray CJL and Evans DB.* Geneva, World Health Organization; 2003:301-318.
4. Weinstein MC, Siegel JE, Gold MR, Kamlet MS, Russell LB: **Recommendations of the Panel on Cost-effectiveness in Health and Medicine.** *JAMA* 1996, **276**:1253-1258.
5. Froberg DG, Kane RL: **Methodology for measuring health-state preferences--II: Scaling methods.** *J Clin Epidemiol* 1989, **42**:459-471.
6. Richardson J: **Cost utility analysis: what should be measured?** *Soc Sci Med* 1994, **39**:7-21.
7. Torrance GW: **Social preferences for health states: an empirical evaluation of three measurement techniques.** *Socio-Economic Planning Sciences* 1976, **10**:129-136.
8. Torrance GW: **Utility approach to measuring health-related quality of life.** *J Chronic Dis* 1987, **40**:593-603.
9. Krabbe PF, Essink-Bot ML, Bonsel GJ: **The comparability and reliability of five health-state valuation methods.** *Soc Sci Med* 1997, **45**:1641-1652.
10. Nord E: **The person-trade-off approach to valuing health care programs.** *Med Decis Making* 1995, **15**:201-208.
11. Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC: **Preferences for health outcomes. Comparison of assessment methods.** *Med Decis Making* 1984, **4**:315-329.
12. Torrance GW: **Measurement of health state utilities for economic appraisal.** *J Health Econ* 1986, **5**:1-30.
13. Martin AJ, Glasziou PP, Simes RJ, Lumley T: **A comparison of standard gamble, time trade-off, and adjusted time trade-off scores.** *Int J Technol Assess Health Care* 2000, **16**:137-147.
14. Salomon JA, Murray CJL: **A multi-method approach to measuring health state valuations.** *Health Econ* 2004 in press.
15. Dolan P, Gudex C, Kind P, Williams A: **The time trade-off method: results from a general population study.** *Health Econ* 1996, **5**:141-154.
16. Brazier J, Roberts J, Deverill M: **The estimation of a preference-based measure of health from the SF-36.** *J Health Econ* 2002, **21**:271-292.
17. Fryback DG, Dasbach EJ, Klein R, Klein BE, Dorn N, Peterson K, Martin PA: **The Beaver Dam Health Outcomes Study: initial catalog of health-state quality factors.** *Med Decis Making* 1993, **13**:89-102.
18. Louviere JJ, Hensher DA, Swait JD: **Stated choice methods: analysis and application.** *Cambridge, Cambridge University Press* 2000.
19. Koop G, Poirier DJ: **Rank-ordered logit models: an empirical analysis of Ontario voter preferences.** *Journal of Applied Econometrics* 1994, **9**:369-388.
20. Beggs S, Cardell S, Hausman J: **Assessing the potential demand for electric cars.** *Journal of Econometrics* 1981, **16**:1-19.
21. Adamowicz W, Louviere J, Swait J: **Combining stated and revealed preference methods for valuing environmental amenities.** *Journal of Environmental Economics and Management* 1994, **26**:65-84.
22. Thurstone LL: **A law of comparative judgment.** *Psychological Review* 1927, **34**:273-286.
23. Ryan M, Farrar S: **Using conjoint analysis to elicit preferences for health care.** *BMJ* 2000, **320**:1530-1533.
24. Bosch JL, Hammitt JK, Weinstein MC, Hunink MG: **Estimating general-population utilities using one binary-gamble question per respondent.** *Med Decis Making* 1998, **18**:381-390.
25. Johannesson M, Jonsson B, Borgquist L: **Willingness to pay for antihypertensive therapy--results of a Swedish pilot study.** *J Health Econ* 1991, **10**:461-473.
26. Cutler DM, Richardson E: **Measuring the health of the U.S. population.** *Brookings Papers on Economic Activity. Microeconomics* 1997:217-271.
27. Erens B: *Health-related quality of life: general population survey. London, Social and Community Planning Research*; 1994.
28. **Health State Valuations from the British General Public, 1993 [data file, SN3444].** *Colchester, Essex, The Data Archive*; 1995.
29. Dolan P: **Effect of age on health state valuations.** *J Health Serv Res Policy* 2000, **5**:17-21.
30. Dolan P: **Modeling valuations for EuroQol health states.** *Med Care* 1997, **35**:1095-1108.
31. Kind P, Dolan P, Gudex C, Williams A: **Variations in population health status: results from a United Kingdom national questionnaire survey.** *BMJ* 1998, **316**:736-741.
32. Gudex C, Dolan P, Kind P, Williams A: **Health state valuations from the general public using the visual analogue scale.** *Qual Life Res* 1996, **5**:521-531.
33. Dolan P, Roberts J: **Modelling valuations for Eq-5d health states: an alternative model using differences in valuations.** *Med Care* 2002, **40**:442-446.
34. Rabin R, de Charro F: **EQ-5D: a measure of health status from the EuroQol Group.** *Ann Med* 2001, **33**:337-343.
35. Luce RD: *Individual choice behavior: a theoretical analysis. New York, John Wiley & Sons, Inc.*; 1959.
36. McFadden D: **Conditional logit analysis of qualitative choice behavior.** In: *Frontiers in econometrics.* Edited by: *Zarembka P.* New York, Academic Press; 1974:105-142.
37. Chapman RG, Staelin R: **Exploiting rank ordered choice set data within the stochastic utility model.** *Journal of Marketing Research* 1982, **19**:288-301.
38. Allison PD, Christakis NA: **Logit models for sets of ranked items.** *Sociological Methodology* 1994, **24**:199-228.
39. Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q: **Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2.** *Med Care* 1996, **34**:702-722.
40. Brazier J, Usherwood T, Harper R, Thomas K: **Deriving a preference-based single index from the UK SF-36 Health Survey.** *J Clin Epidemiol* 1998, **51**:1115-1128.
41. Busschbach JJ, McDonnell J, Essink-Bot ML, van Hout BA: **Estimating parametric relationships between health description and health valuation with an application to the EuroQol EQ-5D.** *J Health Econ* 1999, **18**:551-571.
42. Stevens SS: **On the theory of scales of measurement.** *Science* 1946, **103**:677-680.

43. Yellott JI: **The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution.** *Journal of Mathematical Psychology* 1977, **15:**109-144.
44. Critchlow DE, Fligner MA, Verducci J: **Probability models on rankings.** *Journal of Mathematical Psychology* 1991, **35:**294-318.