

Rescaling quality of life tariffs from discrete choice experiments for use as QALYs: a cautionary tale

Terry N Flynn¹, Jordan J Louviere², Anthony AJ Marley³, Joanna Coast^{4*}, Tim J Peters⁵

¹Department of Social Medicine, University of Bristol, Canynge Hall, Whiteladies Road, Bristol BS8 2PR, UK

²Centre for the Study of Choice, University of Technology Sydney, City - Haymarket Campus, Broadway NSW 2007, Sydney, Australia

³ Department of Psychology, University of Victoria, P.O. Box 3050, Victoria, B.C. V8W 3P5, Canada

⁴ Department of Health Economics, Public Health Building, University of Birmingham, Birmingham B15 2TT, UK

⁵ Department of Community Based Medicine, Department of Community Based Medicine, University of Bristol, 25 Belgrave Road, Bristol BS8 2AA, UK

*Corresponding author

Email addresses:

TNF: terry.flynn@bristol.ac.uk

JJL: jordan.louviere@uts.edu.au

AAJM: ajmarley@uvic.ca

JC: j.coast@bham.ac.uk

TJP: tim.peters@bristol.ac.uk

Abstract

Background

Researchers are increasingly investigating the potential for ordinal tasks such as ranking and discrete choice experiments to estimate QALY weights. However, the assumptions of random utility theory, which underpin the statistical models used to provide these estimates, have received insufficient attention. In particular, the assumptions made about the decisions between living states and the death state are not satisfied, at least for some people. Estimated QALY tariffs are likely to be incorrectly anchored with respect to death (zero) in such circumstances.

Methods

Data from the ICECAP valuation exercise were analysed using an ordinal model proposed previously to estimate QALY-like tariffs. Sensitivity analyses were conducted to vary the proportion of people who conformed to the random utility model underpinning the analyses.

Results

Only 26% respondents conformed unequivocally to the assumptions of random utility theory. At least 14% of respondents unequivocally violated the assumptions. Varying the relative proportions of conforming respondents in sensitivity analyses led to large changes in the estimated QALY tariffs, particularly for lower-valued states. As a result these tariffs could be either positive (considered to be better than death) or negative (considered to be worse than death).

Conclusions

Use of a statistical model such as conditional (multinomial) regression to estimate QALY tariffs from ordinal data is inappropriate in the presence of respondents who do not conform to the assumptions of random utility theory. This is clearest when estimating tariffs for the large proportion of respondents typically observed in valuation samples who refuse to consider any living state to be worse than death: in such circumstances the model cannot be estimated. Only a valuation task requiring respondents to make choices in which both length and quality of life vary can produce estimates that properly reflect the preferences of all respondents.

Background

The fundamental assumption underlying the quality-adjusted-life-year (QALY) model is that the estimated utilities should reflect the relative desirability of health states [1].

The aim of the QALY approach is to allow comparisons of interventions that affect life expectancy to those that affect quality of life. Several elicitation methods have been proposed to estimate QALY weights ('tariffs' in EQ-5D EuroQol parlance) reflecting people's preferences, with recent interest in methods requiring only ordinal respondent preferences [2-4]. Ranking responses such as stating that A is preferred to B are cognitively easier than stating by *how much* A is preferred to B, and so make less stringent assumptions about the cognitive abilities required to compare options.

Discrete choice experiments (DCEs) are the most common type of ordinal task used in health services research to estimate utilities based on patient choices. DCEs can estimate health state tariffs for two reasons. First, they are compatible with Lancaster's theory of value, which states that the total utility of a good or service can be decomposed into utilities of characteristics that describe it [5]; given an appropriate statistical design [6], utilities of alternative health states (profiles/specifications) allow one to infer utilities of attribute levels that describe them. Second, they are consistent with random utility theory (RUT), a well-tested theory of human decision-making [7,8]. RUT assumes that the total utility of a good/service can be expressed as the sum of two components, one fixed (systematic), and a second random (stochastic). If the random component is an independently and identically distributed (iid) extreme value type 1 (Gumbel) random variate, then the underlying choice process is consistent with McFadden's (1974) conditional (multinomial) logit model, and this model can be used to estimate the elements of the fixed component [7]. That is, the relative choice

frequencies reveal the individuals' preferences (utilities), which can be estimated from the frequencies as a function of attribute levels.

Ranking and best-worst experiments can be viewed as generalisations of DCEs [2-4,9] that can be used to obtain data to estimate utilities of individual attribute levels (and their interactions, given larger designs), or the utility of a profile. Two recent papers recognise and discuss the potential of RUT-based choice tasks to estimate such tariffs. Salomon and McCabe *et al* proposed omitting length of life as a variable in the main valuation task [2,3], and asking respondents to choose between impaired living states and the death state. Specifically, respondents are required to rank several states (profiles) with death as one of the states. The idea is that by including the death state somewhere on the latent variable (continuum of health or quality of life), the difference between any given health (or quality of life) state and death on this variable can be estimated from the probability of choosing death over a particular (usually very bad) state. This probability can be estimated at an individual level (if the respondent has made repeated choices) or at the sample level using the proportion of people choosing death.

This paper explores the model proposed by these authors using data and results from the ICECAP (Investigating Choice Experiments for the Preferences of Older People – ICEPOP project – CAPability instrument) DCE. The ICECAP measure provides an index of capability for older people. It is not a QALY measure and is not intended to be such. Nevertheless, it is possible to treat the ICECAP data as if they were being used to generate QALYs, and thus to explore this model using the data generated by the valuation exercise for the ICECAP measure. Using these data, this paper shows

that there is no reason to expect that the model estimates reflect the true QALY trade-offs that the respondents will make. In fact, the utility estimates from such models will agree with those from time trade-off (TTO) or standard gamble (SG) tasks only by chance. Thus, the aims of this paper are to:

- 1) estimate QALY-like tariffs from the ICECAP DCE data using the common model proposed by Salomon and McCabe *et al*; and
- 2) illustrate that their model estimates are sensitive to the proportion of people whose choices are consistent with RUT.

The paper concludes with a consequent research agenda.

Methods

Data – the ICECAP index of capability for older people

The ICECAP instrument was designed to give a set of general capability values for the UK population aged 65+ [10]. By focusing on general quality of life rather than health or health-related quality of life the measure can be used to compare across health and social care interventions. The measure has five attributes (attachment, security, role, enjoyment and control), each varied over four levels. An initial set of tariffs was generated using preference-elicitation methods. Development of the attributes is described elsewhere [10] and the population level tariffs are available as a reviewers' appendix.

Sampling was restricted to those aged 65 and over, using the sampling frame of respondents to the Health Survey for England (HSE). The survey was interviewer-administered in respondents' homes. The main HSE survey (conducted 6-12 months earlier) provided additional data, such as: basic socio-demographic information;

health; nature of locality and environment; social support; participation and contact with others; and general well-being.

The main valuation task was a best-worst scaling (BWS) exercise [11,12] that focuses on ‘profiles’ of ‘attributes’; that is, respondents choose their most and least preferred attribute levels in each profile (quality of life state) they evaluate [4]. This minimises cognitive burden for respondents in this age group, important because traditional DCE tasks requiring the comparison of entire quality of life states can be cognitively difficult. However, two (simple) comparisons were included: 1) comparing each state with a respondent’s own quality of life; and 2) comparing each state with death. The DCE with the latter choice included is described below.

Design of the discrete choice experiment

Varying five attributes ($K=5$), each with four levels ($L=4$), meant that the total number of possible states was $4^5=1024$. Due to practical constraints it was not possible to recruit enough respondents to estimate interactions. Therefore, two versions of an orthogonal main effects plan (OMEPE) obtained from this website (<http://www.research.att.com/~njas/oadir/>), as suggested by Street *et al* [13] were administered. This OMEPE was used to make survey version ‘A’, and its foldover (levels 1 and 4, and levels 2 and 3, are swapped) was used to make survey version ‘B’. Respondents were randomly allocated to receive version A or version B, and each version contained 16 quality of life states. For each of the 16 states, respondents were first asked if the quality of life state in question was at least as good as their own life. If they answered yes, it was assumed that they also considered the state to be better than immediate death. This assumption that nobody would rather die than

continue living may be questionable but could not be tested, given concerns expressed by the ethics committee. If they answered no, they were asked if they considered the state in question to be 'a life worth living'.

Random utility theory

As previously noted, RUT assumes that the utility of each state/profile has a fixed (systematic) component and a random (stochastic) component. Utilities are not known with certainty by researchers, hence are random variables. Thus, the probability that state i is chosen as best is equal to the probability that its utility is greater than the utilities of all other states in a particular choice set. Respondents are assumed to try to choose the state with the greatest utility. To operationalise the model one must make assumptions about the probability distribution of the random utility component.

The conditional logit statistical model

The McFadden multinomial representation of random utility theory (also set out by Holman and Marley in Luce and Suppes [14]) assumes that each random utility component is distributed as an iid extreme value type 1 (EV1 or gumbel) random variate with zero mean and fixed variance [7,15]. This results in the distribution of the difference between two states being logistic [16]. Thus, the conditional logit model assumes that each respondent perceives the difference between state 11111 and death as a random variable from a logistic distribution. McCabe *et al* state that the odds of state j being chosen over state k is $\exp\{\mu_j - \mu_k\}$, where μ_i is the utility of state i [3], with the log odds estimating the utility difference of health states j and k . However, as is true of all limited (discrete value) dependent variable models, this model has an identification problem due to the error variance, or equivalently, the scale of the utility estimates being confounded with the model parameters. The importance of this is explained in the next section.

Model estimation

The ICECAP DCE essentially asked respondents to choose between each state and death (subject to the assumption that a respondent would not prefer to die immediately); hence, one could use a regression model to estimate tariffs that assign a zero value to death based on people's preferences, as proposed by McCabe *et al* and Salomon [2,3]. However, such model estimates from the ICECAP DCE are likely to be unreliable because the proportion of people who choose death as most preferred should tend to zero as quality of life states improve, yielding much less precise estimates. Indeed, if all respondents in a sample agreed that (for example) all states at least as good as 33333 (a state where each attribute has level 3) were worth living, one could not estimate the additional utility of level 4 compared with level 3 for any attribute. Another issue is that the assumption of constant error variance within and between respondents in the DCE is unlikely to be true because people are likely to be more consistent in their preferences as the quality of life state on offer becomes more or less attractive. For all these reasons, choice consistency is likely to be higher for very attractive or unattractive states, and lower for states in between.

To minimise these problems, the full set of anchored tariffs was not estimated directly from the DCE. Instead, the final tariffs from Coast *et al* (unpublished data – see reviewer's attachment) – which are population-level best-worst estimates of quality of life anchored such that state 11111 has zero utility – were rescaled using the DCE estimates of two quality of life states. The importance of using a minimum of two states from the DCE to rescale is clear from consideration of the random utility component. As noted earlier, estimates in all limited dependent variable models are

confounded with the error variance [17]; thus, the estimate of the utility of (for example) state 11111 relative to death is actually its utility *divided by the standard deviation of the error distribution*. So, any particular odds ratio (for 11111 relative to death) is consistent with a large (small) difference in utility divided by a large (small) error standard deviation or, indeed, an infinite number of difference-standard deviation combinations. So, one must correct for this confound before rescaling BWS tariffs, which can be done in the same way that one calculates willingness to pay from a DCE [17]. That is, one can divide $k-1$ utility estimates by the k -th estimate to render the estimates scale-free.

The paragraph above explains why estimates from *at least* two states must be used to rescale the BWS tariffs. However, to maximise precision and minimise the chances of aggregating across choices with different error variances, it makes sense to choose *only* two states that are close to death to maximise the number of respondents who potentially will trade with death. So, the estimates of states 22222 and 11111 relative to death were used to do this. The estimate of state 22222 relative to death divided by that of state 11111 relative to death is scale-free; this ratio was used to rescale the BWS tariffs so that zero represents the death state (rather than zero representing state 11111 as in Coast *et al*). The statistical model is set out more formally below.

Rescaling to ensure death has zero utility

For individual i the utility of state j over that of state 11111 is given by:

$$U_{ij} = \alpha D_j + \vec{\beta} \vec{X}_j + \varepsilon_{ij} \quad (1)$$

\vec{X}_j is a vector of (5 x 3=15) dummy variables, each one representing (for quality of life state j) the additional utility of the level of a given attribute over that of level 1 and $\vec{\beta}$ is the vector of coefficients of the dummy variables. D_j is a dummy variable

equal to 1 for the death state, 0 otherwise. Only the estimated coefficient on the death dummy variable ($\hat{\alpha}$) and estimated coefficients on the five dummy variables representing the additional utility of level 2 over level 1 for each attribute are of interest (State 11111 has zero utility in the model which is estimated in Stata using the *clogit* command.) Thus, $tariff_{22222}$, the tariff of state 22222, is equal to the sum of the five associated elements of $\vec{\beta}$. Furthermore, *ratio*, the base-case scale-free utility of state 22222 relative to state 11111 is:

$$\frac{tariff_{22222} - \hat{\alpha}}{tariff_{11111} - \hat{\alpha}} = ratio$$

Solving for $\hat{\alpha}$:

$$\frac{tariff_{22222} - ratio \cdot tariff_{11111}}{1 - ratio} = \hat{\alpha} \tag{2}$$

Substituting $tariff'_{11111}$ and $tariff'_{22222}$ (the estimated tariffs from the BWS valuation exercise) into (2) along with the value of *ratio* from the DCE gives $\hat{\alpha}'$ (the estimated tariff for death in the BWS model).

Following the rescaling procedure in Coast *et al* of subtracting 1/5 of $tariff_{death}$ from all the BWS attribute level utilities ensured that the ‘bottom’ anchor (here, death) had zero utility. Dividing by the resulting tariff of state 44444 ensured that the ‘top’ anchor (44444) had utility of one. This model assumes all respondents conform to random utility theory in making choices between living states and death. However, as is discussed next, there are reasons to suspect that this may be incorrect in the context of the death state.

Decision-making processes used by respondents when considering death

Consider a case where a respondent makes a statement along the lines of “life is always worth living”. This implies that for that respondent no quality of life state is worse than (or even equal to) death. Thus, those who consider all life worth living, choose between living and death states deterministically instead of stochastically: the probability of choosing a living state as preferred is identically equal to one. Under the random utility model, it can be shown that the probability of choosing a living state over the death state is one if and only if the difference in utility between every living state and death is infinite (see appendix for mathematical proof). Thus, in this case, the death state cannot be assigned any finite value on the latent continuum. Moreover, if one uses conditional logit to model these choices, one implicitly assumes that these people (‘non-traders’) form the right tail of the error distribution. Yet, no distribution exists for them, and the parameter estimate for ‘11111 minus death’ on the latent utility scale simply reflects the relative proportions of people who do and do not trade with death. To show the sensitivity of rescaled tariffs to these proportions, they were varied systematically in simulations using a modified bootstrap procedure [18].

Sensitivity analysis

To show how a random utility model can result in misleading inferences, two hypothetical types of people were considered: 1) people who are traders; and 2) people consistent with the random utility model when trading-off attributes of quality of life, but who are deterministic when comparing states with death (that is, their choice rule is that life is always preferred to death). The proportion of people of each type was varied in a series of analyses by resampling from each of the two types: those who traded with death and those who never traded with death. It should be noted that increasing the proportion of people who traded with death relaxes the

assumption that all people in ICECAP observed not to trade would *never* have traded (and were therefore type 2). Respondents were (re)sampled (with replacement) from the actual samples observed in ICECAP such that the observed frequencies and counts for the two types were as in Table 1 below.

<Insert Table 1 here>

The overall sample size was fixed at that number with complete data in the DCE (282). 50 bootstrap resamples were used for each set of proportions; additional resamples are not required to estimate a bootstrap mean compared with estimating, say, a percentile confidence interval around a mean. The mean of the 50 resamples was calculated for the death dummy variable, which represented the estimated utility of death relative to the omitted state (11111), confounded with the unobserved error variance. The sum of the five dummy variables (one for level 2 of each attribute) represents the utility estimate for state 22222 relative to state 11111, again confounded with the unobserved error variance. Equations (1) and (2) were used to estimate the position of death on the BWS scale, enabling rescaled tariffs to be constructed, which make the same assumptions as those in QALYs.

Results

Data were collected between October 2005 and January 2006, with 478 individuals sampled. This yielded 315 (66%) fully productive interviews (respondents reached the end of the interview), of which 282 answered all 16 DCE comparisons with death. Of the 282 respondents with complete DCE data, 73 (25.6%) considered at least one state to be worse than death. For survey version A, 46 of 151 (30.5%) were willing to trade

with death, and for version B, 27 of 131 (20.6%) were willing to trade. The higher version A percent likely reflects the fact that state 11111 only appears in A; the worst possible state in version B (state 41111 according to the BWS tariffs) had “attachment” with a better level than 1.

ICECAP base-case rescaled tariffs

Table 2 gives:

- BWS tariffs from Coast *et al* representing quality of life states scaled such that state 11111 has value of zero and state 44444 has value of one,
- Raw DCE (‘part-worth’) estimates, together with tariffs for key states 22222 and death (both with respect to state 11111) and
- BWS tariffs rescaled by the two DCE estimates such that death has utility of zero and state 44444 has utility of one.

<Insert Table 2 here>

Table 2 shows that the DCE estimates are not always rationally ordered for level 2 and above for certain attributes. Indeed, the only estimates that monotonically increase with levels are for “control”, which reflects poor precision due to small numbers who trade with death. As expected, the estimate of the death state is negative because on average it was considered worse than state 11111 (the omitted state in the DCE); only 26% of people were willing to consider at least one living state to be not worth living – that is, 26% were “traders”. Thus, on the latent utility scale state, 22222 is $(2.942 - -0.993)/(0 - -0.993) = 3.964$ times as far from death as state 11111. The rescaling used ensures that the new BWS tariffs in the final column retain this relationship between states 22222 and 11111.

74% of respondents never traded with the death state. However, some of these may have traded on another occasion, or if different quality of life states had been presented. 39 respondents (14%) are recorded as having spontaneously made statements along the lines of “life is always worth living” (often that is was “God-given”) which shows them clearly to be type 2, making choices between living and death states deterministically, not stochastically. There may have been others who made similar comments not recorded by the interviewer and as respondents were not directly asked this question, this 14% can be regarded as the minimum proportion with this view. Moreover, it seems unlikely that the true percentage of people in the UK population being non-traders is as low as this given findings from the main EQ-5D valuation exercise that, for every single living state, a majority of people rated it more highly than immediate death [19]. Therefore, the effect of varying the percentage of people trading with death was investigated in the sensitivity analyses.

Sensitivity analysis

Table 3 presents seven sets of rescaled tariffs. The first represents the base case state in Table 1 using the actual DCE data from the ICECAP valuation survey. Each of the other six sets presents:

- Rescaled BWS tariff for state 11111;
- Rescaled BWS tariff for state 22222 (calculated from the sum of the five level 2 dummy variables); and
- Best-Worst tariffs rescaled using the DCE estimates according to the model presented above.

Each set of results represents the mean of 50 stratified bootstrap replications using sampling probabilities to produce the frequencies in Table 1.

<Insert Table 3 here>

Table 3 shows that as the percent of people considering states worse than death grows (type 1), the rescaled BWS tariffs are ‘stretched’ downwards. Estimated tariffs for unattractive states become less positive, with state 11111 being zero when the estimated coefficient for the death dummy equals zero, implying that the odds of choosing death over any living state is approximately one. As the proportion of type 1 people grows larger than 50%, the estimated tariff of state 11111 becomes negative. Thus, for set 6, which has percentages of traders and non-traders approximately in reverse to what was observed in reality, the mean estimate of 11111 is highly negative. This implies that on average people consider this state to be worse than death.

Discussion

It is important to note that people who refuse to trade with death may still consider an impaired health state to be ‘worth’ a fraction of full health in standard gamble or time trade-off tasks; a person doesn’t have to consider states that occur for sure, in other words are *certain*, to be worse (or better) than death to be indifferent between the *lives/gambles* involving impaired health. A given log odds ratio in the McCabe *et al* model cannot be interpreted as a mean difference in latent utility. Instead, it is an average (with unknown weights) of (at least) two groups of people:

- For one group, there is a distribution of utility differences between 11111 and death. Choices of people in this group conform to RUT with death somewhere on the utility scale (other than minus infinity), and may conceivably vary in repeated samples.

- A second group whose choices are deterministic because they consider all life to be worth living – effectively, for members of this group the utility of death is minus infinity.

The reason why the weights are unknown is that respondents who never trade with death in a given valuation exercise may have traded on another occasion or when faced with a worse health/quality of life state. Furthermore, even if the weights were known, one cannot estimate the ‘true’ population average TTO or SG value because the TTO/SG values for non-traders cannot be recovered in the analysis. The latter point is most obvious when a sample contains only people who always consider life worth living, no matter how bad. The odds that such people will choose death is exactly zero (because crucially, this is *not* a sampling zero), and the utility difference is infinite for a RUT model with EV1 errors (see the mathematical Appendix). Thus, the model proposed by Salomon and by McCabe *et al* cannot be estimated because no-one will choose death, leading to lack of identification of the death dummy variable. In turn, this will lead to all estimates being measured with respect to a remaining state such as 11111. If a given DCE (or ranking/BWS) model purports to estimate tariffs reflecting quantity-quality trade-offs like QALYs, it also must apply to people who never consider health states to be worse than death. The model of Salomon and McCabe *et al* clearly does not satisfy these requirements [2,3].

Limitations

The proportion of people unwilling to consider any ICECAP state to be worse than death in the valuation sample was 0.74. It is unlikely that all were type 2 people, so some of them might with repeated sampling decide that death was better than 11111 (at least, and possibly other states higher on the latent continuum). It also may be that

11111 was insufficiently unattractive for them to trade with death and/or there may be states imaginable to them for which death would be preferable. Nevertheless, as state 11111 was described as having ‘none’ of any of the five (intended as, in some sense, ‘fundamental’) attributes of quality of life, it seems reasonable that most of these respondents cannot conceive of a living state that is both worse than 11111 and worse than death.

Potential variation in rescaled tariffs was shown, but the actual degree of bias (deviation from TTO/SG values) could not be calculated as TTO or SG questions were not asked. The ICECAP study will be repeated with a general population sample that will include at least one TTO question to inform this issue. More generally, questions such as “does this model produce estimates with an acceptably low bias?” will in any case require a TTO/SG estimate of state 11111 (or another state), in which case using a RUM to anchor estimates becomes moot.

Although the RUM used is the same as McCabe *et al* and Salomon, model estimation required more complex analysis to synthesise two different choice processes.

Therefore the approach adopted here leads to less clear inferences. However, the approach is justified due to the cognitive burden that would have been imposed by a traditional DCE and the imprecision in model estimates resulting from comparisons with death.

As the percentage of people willing to trade with death decreased, estimation issues increased because the estimates became very sensitive to the choices of fewer people. This caused problems for bootstrapping because a given bootstrap sample was likely

to include only people with the same preferences, leading to boundary solutions and a failure of maximum likelihood estimation to converge on a global maximum.

Comparisons with previous work

McCabe *et al*'s model yields an estimated difference between the lowest health state and death that implies an odds of 2.0375:1 for these two states [3]. Thus, approximately 2/3 of people thought the living state preferable to death. Whilst it is possible that greater religiosity led to the higher percentage of non-traders in ICECAP, it has been observed that older people are *more* likely to consider very bad states to be worse than death in EQ-5D data [19]. The ICECAP team's use of best-worst methods to estimate the quality of life tariffs is supported by one of McCabe *et al*'s findings, namely that the latter's models were sensitive only to upper and lower rankings. However, it should be noted that neither Salomon nor McCabe *et al* seemed to consider differences in error variances: in the presence of larger variances around middle rankings the assumption of constant variance artificially reduces the sensitivity of results to these.

Salomon found that the TTO value for the worst possible state was not the same as that from a MNL (random utility) ranking model [2]. This is unsurprising as the estimates will only coincide by chance, and a different proportion of people refusing to consider any states as worse than death may have produced answers that agreed with the TTO estimate.

Future work

The TTO and SG methods require people to choose between health states lasting for defined periods of time. One can argue that respondents in DCEs (ranking and BWS) should also do the same; that is, respondents would choose a complete health description (lasting for a given length of time) that they prefer. Thus, researchers should consider including length of life as an experimental factor in future preference elicitation studies that use ordinal response tasks. Such an approach requires a more complex study design that is beyond the scope of this paper but guidelines for such designs are now readily available [6].

Future work that estimates QALY tariffs should ensure that statistical models used to analyse data are consistent with decision-making processes (models) used by people. McCabe *et al* state “research on the thought processes of individuals undertaking ranking exercises would be a valuable contribution to this field” [3]. In fact, this understates the seriousness of the issue; we clearly need research to ascertain under what circumstances and to what extent choices satisfy RUT assumptions. Indeed, it seems reasonable to think that a respondent may conform to RUT in one context (comparing quality of life), but not in another (comparing life with death). It also may be that there is no (easy) way to avoid asking one or more TTO/SG questions to properly estimate the anchor at death, and one of Salomon’s proposed methods did exactly this [2]. In this event, one should try to minimise the context effects for which some TTO tasks previously were criticised [20]. DCE models incorporating length of life as a variable also deserve investigation, but it is unclear if funding bodies are (yet) willing to finance the larger studies that this would require.

Philosophical and psychological issues around aggregation of preferences over people who consider bad states worse than death and those who consider all life worth living are pertinent. Recent work suggests growing realisation that this needs more thought before another large QALY valuation exercise is conducted [21].

Conclusions

Conditional logit estimates of utility differences between a given living state and death can be heavily influenced by the proportion of people who consider a state to be worse than death. It does not, and cannot, take into account the utility values of living compared to death for people who make choices involving death deterministically. Moreover, the greater the number of such people in a given DCE, the more biased the estimate of the utility difference because the (assumed) logistic distribution is not defined for people who make choices deterministically instead of probabilistically.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TNF contributed to study design, carried out all data analyses and drafted the paper.

AAJM provided substantial conceptual input, wrote the mathematical proof

(Appendix) and participated in redrafting of the paper. JJJ designed the methodology,

advised on study design and analysis and participated in redrafting of the paper. JC

participated in the design and running of the study, and in redrafting of the paper. TJP

contributed statistical input to the study and participated in redrafting of the paper. All

authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank the participants in the ICECAP valuation exercise

along with Lucy Natarajan and Kerry Sproston for their roles in the collection of the

data.

References

1. Drummond M, Sculpher M, Torrance GW, O'Brien BJ, Stoddart GL: *Methods for the economic evaluation of health care programmes*, Third edn. Oxford: Oxford University Press; 2005.
2. Salomon JA: **Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data.** *Population Health Metrics* 2003, **1**.
3. McCabe C, Brazier JE, Gilks P, Tsuchiya A, Roberts J, O'Hagan A *et al.*: **Using rank data to estimate health state utility models.** *J Health Econ* 2006, **25**: 418-431.
4. Flynn TN, Louviere JJ, Peters TJ, Coast J: **Best-Worst Scaling: What it can do for health care research and how to do it.** *J Health Econ* 2007, **26**: 171-189.
5. Lancaster K: **A New Approach To Consumer Theory.** *Journal of Political Economy* 1966, **74**: 132-157.

6. Street DJ, Burgess L: *The Construction of Optimal Stated Choice Experiments: Theory and Methods*. John Wiley & Sons Inc; 2007.
7. McFadden D: **Conditional logit analysis of qualitative choice behavior**. In *Frontiers in Econometrics*. Edited by Zarembka P. New York: Academic Press; 1974:105-142.
8. Thurstone LL: **A law of comparative judgment**. *Psychological Review* 1927, **34**: 273-286.
9. Louviere JJ, Hensher DA, Swait J: *Stated choice methods: analysis and application*. Cambridge: Cambridge University Press; 2000.
10. Grewal I, Lewis J, Flynn TN, Brown J, Bond J, Coast J: **Developing attributes for a generic quality of life measure for older people: Preferences or capabilities?** *Soc Sci Med* 2006, **62**: 1891-1901.
11. Marley AAJ, Louviere JJ: **Some probabilistic models of Best, Worst, and Best-Worst choices**. *J Math Psychol* 2005, **49**: 464-480.
12. Finn A, Louviere JJ: **Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety**. *J Public Policy Mark* 1992, **11**: 12-25.
13. Street DJ, Burgess L, Louviere JJ: **Quick and Easy Choice Sets: Constructing Optimal and Nearly Optimal Stated Choice Experiments**. *Int J Res Mark* 2005, **22**: 459-470.
14. Luce RD, Suppes P: **Preference utility and subjective probability**. In *Handbook of mathematical psychology, volume III*. Edited by Luce RD, Bush RR, Galanter E. New York: Wiley; 1965:249-410.
15. Luce RD: *Individual choice behavior*. New York: John Wiley & Sons; 1959.
16. Ben-Akiva M, Lerman SR: *Discrete choice analysis: theory and application to travel demand*. Cambridge, MA: MIT Press; 1985.
17. Hensher DA, Louviere JJ, Swait J: **Combining sources of preference data**. *J Econometrics* 1999, **89**: 197-221.
18. Efron B, Tibshirani R: *An introduction to the Bootstrap*, 1 edn. New York: Chapman & Hall; 1993.
19. Williams A. The measurement and valuation of health: a chronicle. 136. 1995. York, University of York. Centre for Health Economics Papers. <http://www.york.ac.uk/inst/che/pdf/DP136.pdf>
20. Arnesen T, Trommald M: **Are QALYs based on time trade-off comparable? - A systematic review of TTO methodologies**. *Health Econ* 2005, **14**: 39-53.
21. Robinson A, Spencer A: **Exploring challenges to TTO utilities: valuing states worse than dead**. *Health Econ* 2006, **15**: 393-402.

Tables

Table 1: Sensitivity analysis: proportions and numbers of respondents by type

Set	Type 1		Type 2	
	p	$p \times 282$	$1-p$	$(1-p) \times 282$
Base-case				
1	0.10	28	0.90	254
2	0.25	70	0.75	212
3	0.30	84	0.70	198
4	0.40	112	0.60	170
5	0.50	141	0.50	141
6	0.75	211	0.25	71

p is the proportion of respondents who considered at least one ICECAP state to be worse than death; 282 respondents provided complete DCE data

Table 2: BWS, DCE and rescaled BWS (rBWS) estimates

State		BWS	DCE	rBWS
Death tariff		-	-0.993	0
11111 tariff		0	-	0.158
22222 tariff		0.556	2.942	0.626
Attribute	Level			
Attachment	I can have all of the love and friendship that I want	0.254	1.731	0.245
	I can have a lot of the love and friendship that I want	0.233	0.931	0.227
	I can have a little of the love and friendship that I want	0.134	1.313	0.144
	I cannot have any of the love and friendship that I want	-0.013	-	0.021
Security	I can think about the future without any concern	0.179	0.264	0.182
	I can think about the future with only a little concern	0.107	0.479	0.122
	I can only think about the future with some concern	0.066	0.143	0.087
	I can only think about the future with a lot of concern	0.032	-	0.059
Role	I am able to do all of the things that make me feel valued	0.192	0.472	0.193
	I am able to do many of the things that make me feel valued	0.179	0.551	0.183
	I am able to do a few of the things that make me feel valued	0.130	0.385	0.141
	I am unable to do any of the things that make me feel valued	0.015	-	0.044
Enjoyment	I can have all of the enjoyment and pleasure that I want	0.166	0.419	0.171
	I can have a lot of the enjoyment and pleasure that I want	0.164	0.619	0.170
	I can have a little of the enjoyment and pleasure that I want	0.119	0.453	0.131
	I cannot have any of the enjoyment and pleasure that I want	0.017	-	0.046
Control	I am able to be completely independent	0.209	1.631	0.208
	I am able to be independent in many things	0.185	1.137	0.187
	I am able to be independent in a few things	0.108	0.648	0.122
	I am unable to be at all independent	-0.051	-	-0.012

Table 3: sensitivity analysis results

Set		Base	1*	2**	3	4	5	6
% Type 1			10%	25%	30%	40%	50%	75%
11111 tariff		0.158	0.260	0.159	0.128	0.063	0.007	-0.132
22222 tariff		0.626	0.671	0.627	0.613	0.584	0.559	0.497
Attribute	Level							
Attachment	... all the love and friendship	0.245	0.240	0.245	0.247	0.250	0.253	0.260
	... a lot of the love and friendship	0.227	0.224	0.227	0.228	0.230	0.232	0.237
	... a little of the love and friendship	0.144	0.151	0.145	0.142	0.138	0.134	0.125
	...not any of the love and friendship	0.021	0.043	0.021	0.014	0.001	-0.011	-0.041
Security	think about future without any concern	0.182	0.184	0.182	0.182	0.180	0.179	0.176
	think about the future with only little concern	0.122	0.131	0.122	0.119	0.113	0.108	0.095
	think about the future with some concern	0.087	0.101	0.087	0.083	0.075	0.067	0.048
	think about the future with a lot of concern	0.059	0.076	0.059	0.054	0.043	0.033	0.010
Role	able to do all things make me feel valued	0.193	0.194	0.194	0.193	0.193	0.192	0.191
	able to do many things make me feel valued	0.183	0.185	0.183	0.182	0.181	0.179	0.177
	able to do few of things make me feel valued	0.141	0.148	0.141	0.139	0.134	0.130	0.120
	unable to do any things make me feel valued	0.044	0.063	0.045	0.039	0.027	0.016	-0.009
Enjoyment	can have all enjoyment and pleasure	0.171	0.175	0.171	0.170	0.168	0.166	0.162
	can have lot of enjoyment and pleasure	0.170	0.174	0.170	0.169	0.167	0.165	0.160
	can have little of enjoyment and pleasure	0.131	0.140	0.131	0.129	0.124	0.119	0.108
	cannot have any of enjoyment and pleasure	0.046	0.064	0.046	0.040	0.028	0.018	-0.007
Control	able to be completely independent	0.208	0.207	0.208	0.208	0.209	0.209	0.211
	able to be independent in many things	0.187	0.189	0.187	0.187	0.186	0.185	0.183
	able to be independent in a few things	0.122	0.132	0.122	0.119	0.113	0.108	0.095
	unable to be at all independent	-0.012	0.014	-0.011	-0.019	-0.035	-0.050	-0.084

* Based on only three bootstrap samples: maximum likelihood estimation would not converge on global maximum on iteration 2, 3, or 4 for 8 different seeds for random number generator

** Three bootstrap sets abandoned due to failure to converge on an iteration. Fourth seed chosen successfully allowed a set with 50 bootstrap iterations to be estimated

Appendix:

Summary: If the extreme value random utility version of the (conditional) multinomial logistic (MNL) model is assumed to hold and the choices of a respondent are such that $P(L, D)$ is identically (not statistically) equal to 1 for all living states $L \neq D$, where D is death, then $u(D)$ cannot be assigned any finite value, such as 0.

Let L denote a typical quality of state, i.e. a living state, and D death. Let $P(L, D)$ denote the probability that an individual respondent chooses L over D. Assume that the respondent's choices satisfy the random utility version of MNL. That is, there is a (standard) extreme value (type 1) random variable ε such that $(\Pr(\varepsilon \leq t) = \exp(-e^{-t}); (-\infty < t < \infty))$, and scale values $u(L)$ and $u(D)$ such that

$$P(L, D) = \Pr(u(L) + \varepsilon > u(D) + \varepsilon')$$

where ε and ε' are independent samples of the random variable. Then classic results show that

$$P(L, D) = \frac{e^{u(L)}}{e^{u(L)} + e^{u(D)}} = \frac{e^{u(L)-u(D)}}{e^{u(L)-u(D)} + 1} \quad (1)$$

Case 1: When is $P(L, D) \equiv 1$ for all living states $L \neq D$?

(1) implies this is the case if, and only if, for all $L \neq D$, $u(L) - u(D) = \infty$. The only possible solutions are:

- i) for all $L \neq D$, $u(L) = \infty$, $u(D) < \infty$,
- ii) for all $L \neq D$, $u(L) > -\infty$, $u(D) = -\infty$.

Comment 1:

In case 1i) we would expect the respondent to choose randomly between any two living states, L_1 and L_2 , provided neither was death (though that choice probability would not be given by the RUM MNL as it can be interpreted as given by

$$P(L_1, L_2) = \Pr(\infty + \varepsilon > \infty + \varepsilon') = \frac{\infty}{\infty + \infty},$$

which is undefined. In any case such choices do not seem to be made in real data. So we are left with case 1ii), where the utilities of different living states L_1, L_2 , neither being death, can (but do not have to) differ, and all such states are chosen deterministically (i.e. with probability one) over death.

Comment 2:

We can fit the MNL to the data of a respondent satisfying Case 1ii), but we cannot use death as a 'referent' state that is assigned some finite value, for example, 0, as the true value of death is $-\infty$.

Case 2: When is $P(L, D) > 1 - \delta$, with δ 'small', for all living states $L \neq D$?

Using (1), routine algebra shows that this holds provided that, for all $L \neq D$,

$$u(L) - u(D) > \ln \frac{1 - \delta}{\delta}.$$

However, as δ approaches zero, the right hand side approaches infinity, and thus, for all $L \neq D$, $u(L) - u(D)$ approaches ∞ which brings us back, effectively, to Case 1.

Conclusion: If the extreme value random utility version of the MNL model is assumed to hold and the choices of a respondent are such that $P(L, D)$ is identically (not statistically) equal to 1 for all living states $L \neq D$, where D is death, then $u(D)$ cannot be assigned any finite value, such as 0.

Additional files provided with this submission:

Additional file 1: tariffspaper.pdf, 243K

<http://www.pophealthmetrics.com/imedia/2143979639180058/supp1.pdf>