The use of risk-based discrete choice experiments to capture preferences over health states.

Angela Robinson\textsuperscript{1}, Anne Spencer\textsuperscript{2} & Peter Moffatt\textsuperscript{1}

\textsuperscript{1} University of East Anglia
\textsuperscript{2} University of Exeter Medical School.

Abstract

We report the results of an exploratory study that set out to develop a risk-based DCE approach to derive utility values for health states. The design allows states to be anchored to normal health and death allowing utility values and probability weights to be derived directly within the DCE. It also allows worse than dead states to be valued in the same manner as better than dead states. The DCE results are compared to those derived from a modified SG, where risk is present on both sides, under expected utility and rank dependent utility. In both DCE and modified SG, adjusting for probability weighting in rank dependent utility significantly reduces the values. We discuss the implications of our findings for future applications.

Word count:9064
The use of risk-based discrete choice experiments to capture preferences over health states.

Angela Robinson¹, Anne Spencer² & Peter Moffatt¹

¹ University of East Anglia
² University of Exeter Medical School.

1. INTRODUCTION

Discrete choice experiments (DCEs) allow a number of characteristics to be traded off against one another and are becoming increasingly popular in health economics. Although the origins of the DCE approach lie in marketing, they were later applied to the valuation of aspects of health care not easily captured using conventional Quality of life measures, such as the type of services received (see, for example, Damman et al, 2012 and Scott et al, 2003). There has, however, been recent interest in using the DCE method to derive health state utilities for use in QALY calculations and there remains uncertainty about how that is best done (Stolk et al, 2010). This paper adds to the small literature on how utility values for health may be derived directly within a DCE.

The motivation for the use of DCEs to elicit utility values is often linked to well-known problems that exist with more traditional value elicitation techniques such as Standard Gamble (SG) and Time Trade Off (TTO). For example, the SG has been criticized on the basis that Expected Utility (EU) is not a descriptively valid theory of decision making under risk. Furthermore techniques (such as SG and TTO) that set out to elicit an individual’s point of indifference are regarded as more cognitively demanding than those involving pair-wise choices (Ratcliffe et al, 2011).

There are a number of methodological challenges, however, to be addressed in deriving health state utility values in particular within a DCE. In more traditional applications of DCEs attributes can reasonably be assumed to contribute to utility in an additive way, albeit with the possibility of interaction effects between attributes. Moving towards applications such as health state utility assessment typically requires the inclusion of attributes such as risk (van Houtven et al, 2011), duration (Bansback et al, 2012) and numbers treated (Robinson et al, 2010). This presents different challenges than have traditionally arisen in DCEs in terms of the appropriate functional form of the model. In particular, the need to combine such attributes in a multiplicative, rather than additive, manner (Flynn, 2010).

Estimating utility values for health states within a DCE requires health states to be anchored to normal health (generally assigned a value of ‘1’) and dead (generally assigned a value of ‘zero’). There are DCE studies that look at comparison of health states, without trying to link them to a normal health/dead scale (for example, Hakim et al, 1999), but the results cannot then be used as utility values and incorporated into QALY calculations. Alternatively, researchers have used techniques such as SG or TTO alongside the DCE in order to provide anchors to normal health and death (Stolk
et al, 2010). But as the authors themselves point out, if DCE is being used in order to overcome perceived problems with SG and TTO, such ‘solutions’ are obviously problematic.

Recent studies have used a ‘TTO-like’ format and have linked health states to normal health and dead within a DCE by either including ‘dead’ as a state or survival duration as an attribute (Brazier et al, forthcoming and Bansback et al, 2012). The approach, termed $\text{DCE}_{\text{TTO}}$, includes health states and survival duration (not including zero) as attributes, but does not include immediate death as an option. For example, Bansback and colleagues use this approach to value a range of EQ-5D states (Bansback et al, 2012). In the $\text{DCE}_{\text{TTO}}$, normal health is set at one and values worse than dead can be inferred ‘indirectly’ at a sample level, by using the coefficients on attribute levels as incremental reductions in quality of life. Thus, setting normal health equal to 1 and subtracting incremental decreases in quality of life associated with attribute levels, there will come a point when the values lie below zero. Values inferred less than zero are then taken to signify that the state is worse than dead. One possible drawback with this approach is that there is no independent verification that any respondent considers that state to be worse than dead.

Another way in which utility values may be linked directly to death within a DCE is to include some risk of immediate death as an option. One advantage of using risky choices is that there is a body of research looking at decision making under risk (see, for example, Karni, 2009; Machina 1988) and there has been some success in adjusting for biases in risky choices (Doctor et al, 2010; Abellan-Perpinan et al, 2009). Including risk as an attribute in a DCE will necessarily involve presenting respondents with two risky treatments, an approach that has been used successfully in other studies (McCord and de Neufville 1986 and Carthy et al, 1998). Whilst traditional SG approaches generally involve the certainty of the ‘target’ health state, we use a ‘modified’ SG-type approach here to denote that risk appears in both options.

If we are to use risk-based DCE, it is important to consider how the theory of random utility might be adapted to take on board the recent advances in decision-making under risk. The random utility theory of McFadden and Heckman underpins the analysis of DCEs. It models decision making as a stochastic process around expected utility. In contrast, there are a number of non-expected utility models of decision-making. For example, Rank Dependent Utility (RDU) assumes that people over- or under-weight probability and so it incorporates a probability weighting function in its specification of decision making (Quiggin, 1993). Cumulative Prospect Theory assumes a probability weighting function and also allows for people to experience greater changes in utility from losses compared to gains (Tversky and Kahneman, 1992).

An important step forward has been made by de Palma et al (2008) who argued that these non-expected utility functions can also be used to underpin random utility theory. They outline the types of data needed to infer these models. One important paper that has looked into the use of non-expected utility functions in DCE models looks at the treatment and risky side-effects of Crohn’s disease (van Houtven et al, 2011). The authors found evidence of non-linearity in how these risks were perceived and derived lower utility values under assumptions of RDU than EU. This led them to
argue that traditional SG methods assuming EU are biased but stressed that it was difficult to do a direct comparison as the nature of the risks included in their DCE was very different to those commonly used in SG. The authors call for more research into the use of DCEs using risks that are more typically included in SG (van Houtven et al, 2011). One recent study (Ratcliffe et al, 2011) compares SG (and TTO) to a ‘best-worst scaling’ DCE but we know of no direct comparison of SG with a ‘standard’ DCE approach.

One criticism of traditional value elicitation techniques, such as SG and TTO is that the procedures for valuing states worse than dead involve a fundamental departure from those used to value better than dead states. Given the large body of evidence showing that responses can be affected by descriptive and procedural invariance (Tversky et al, 1988) we argued previously that such evidence must call into question the validity of aggregating better than and worse than dead scores generated by two different procedures (Robinson and Spencer, 2006). It can be shown, however, that a technique which presents respondents with choices over two risky treatments allows states worse-than-dead to be valued in the same manner as better than dead states.

We set out here to develop a method whereby utility values may be derived within a DCE that anchors them directly to normal health and death. Further, we set out to allow worse than dead health states to be derived in the same manner as better than dead states. We compare that approach with a ‘modified’ SG.

The aims of this research are therefore:

1) To develop a method for eliciting values for health states, anchored to normal health and dead, within a risk-based DCE.
2) To develop a framework in which values for ‘better than dead’ and ‘worse than dead’ health states can be elicited in the same manner.
3) To compare EU and non-EU models of risky choice behaviour within a DCE.
4) To compare the results of the DCE model(s) with the modified SG.

2. METHODS

2.1 Overview of the survey

There were 60 participants recruited from the population of second and third year students studying Economics or Geography at the Universities of London (Queen Mary) and Exeter in 2011/12. Data were collected by means of small groups comprising on average between 8 and 9 participants. Groups were generally convened by two authors (AS and AR) although it was not possible in all cases. Respondents were invited to take part either through e-mail (at Queen Mary) or through the experimental laboratory (FEELE at Exeter University). All subjects were paid £10 for taking part.

The groups began with a brief introduction to the aims of the study and the questionnaire then aimed to elicit values for three EQ-5D health states (21121, 22222 and 22323). The first part of the questionnaire asked respondents to rank the health states that were presented on small cards along with normal health (11111) and ‘immediate death’. This was followed by DCE questions (15) and modified SG questions (3). The order in which the DCE and modified SG questions appeared was
randomised. Finally respondents answered a series of 4 questions designed to elicit risk attitudes using money lotteries.

2.2 The DCE questions
In the DCE part of the questionnaire, a series of questions presented respondents with two risky treatments, labeled A and B. All risky treatments involved some chance \( p \) of an outcome (21121, 22222, 22323, or immediate death) and an associated chance, \( 1-p \), of normal health (11111). Thus, normal health appeared in all treatments. A typical question is shown in Figure 1 and used graphical displays to illustrate risk information. Treatment A offers a 10% chance of normal health and a corresponding 90% chance of health state 21121. Treatment B offers a 99% chance of normal health and 1% chance of immediate death. We simplify this notation henceforth as Treatment A offers a 90% chance of 21121 and Treatment B offers a 1% chance of death. It is important, however, not to lose sight of the fact that there is always an associated chance of normal health.

Respondents were asked to suppose that they had some condition and they were faced with two different treatments for that condition. They were asked to tick one of three possible responses, namely: prefer A; equally preferable, prefer B. We elected to include the ‘indifference’ option in the choice data as we wanted to maximize the similarities across the DCE and modified SG approaches. By having risk on both sides, we hoped to overcome the ‘certainty effect’ bias observed in other studies (van Osch and Stiggelbout, 2008; van Osch et al, 2006 and Hershey and Schoemaker, 1985). By allowing an ‘equally preferable’ response we hoped to avoid ‘forcing’ a preference as it has been noted that may not always be appropriate (see Viney et al, 2002).

Whilst we have used the EQ-5D descriptive system for convenience, it is important to stress here that we are not setting out here to derive an alternative set of weights for that system. As developing and demonstrating a methodology is our aim here, we opted for a very simple design involving only two attributes- outcome and risk. The DCE questions varied on one or more of the two attributes shown below:

- The outcome (health states 21121, 22222, 22323 or immediate death) coloured yellow, green, grey and blue respectively.
- The probability of that outcome (1%, 5%, 10%, 20%, 30%, 40%, 50%, 70%, 90%).

The attributes and levels set in this study produced a total of 630 different choices. We chose to include all non-dominated choices in this exploratory study. The main rationale for this decision was the uncertainty surrounding use of optimal design for multiplicative model (Flynn, 2010) and criticisms of previous studies that have used fractional factorial designs unnecessarily limiting the scope of the analysis (Viney et al, 2002).

1 The total number of scenarios was \( 9 \times 1 \times 4 = 36 \). The number of ways of choosing \( r=2 \) scenarios at random from \( n=36 \) is \( \frac{n!}{(n-r)!r!} = 36.35/2 = 630 \).
There were three types of dominance that arose in the study due to the levels of risk, health states or both. Choices could be ‘risk-dominated’ in that they involved the same health state but a different level of risk attached to that state. For example, suppose that Treatment A offered a 40% chance of health state 21121 (and associated 60% chance of normal health) and Treatment B offered a 30% chance of the same health state (and associated 70% chance of normal health). Treatment B clearly dominates Treatment A in this case. We elected to ask all participants (a different) one of the 144 ‘risk-dominated’ choices contained in the full factorial, as a simple test of consistency, and so included 60 risk-dominated choices.

As there is a ‘logical’ ordering of health states in that 21121 > 22222 > 22323, there is another type of dominance that we term ‘state-dominated’. For example, suppose that Treatment A offered a 40% chance of health state 21121 (and associated 60% chance of normal health) and Treatment B offered a 40% chance of 22222 (and associated 60% chance of normal health). Treatment A clearly dominates Treatment B in this case as 21121 is strictly better than 22222. The full factorial contained a total of 27 such choices which we randomly allocated. Finally, choices could be ‘risk/state dominated’ in that they involved a lower risk of a less severe state. The full factorial contained a total of 108 such comparisons which we retained and which were randomly allocated.

Respondents were presented with a set of 15 DCE questions. The first question was one drawn randomly from the 144 ‘risk dominated’ comparisons described above. Each respondent was presented with a further 8 questions drawn randomly from the full factorial design. In addition, respondents were presented with a common set of 6 questions that were interspersed with those randomly allocated. The 6 ‘common’ questions were a series that set out to allow the utility value of one health state-22222- to be determined at the level of the individual respondent (or for at least allow a range to be determined). Table 1 outlines these six pair-wise comparisons with the question number denoting the order they appeared in the questionnaire. The last column of Table 1 calculates the utility value for state 22222 that would be derived if respondents were indifferent between Treatments A and B (assuming EU preferences).

---

2 For example, suppose that Treatment A offered a 30% chance of health state 21121 (and associated 70% chance of normal health) and Treatment B offered a 40% chance of 22222 (and associated 60% chance of normal health). Treatment A clearly dominates Treatment B, as is offers a lower risk of less severe illness.

3 In total our design included 546 choices (630-144=546): 351 non-dominated and 195 dominated choices that we elected to retain (60+27+108=195).
Table 1: The 6 DCE questions answered by all respondents

<table>
<thead>
<tr>
<th>Question</th>
<th>Risk of 22222* in Treatment A</th>
<th>Risk of death* in Treatment B</th>
<th>Implied value of 22222 under EU</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>90%</td>
<td>5%</td>
<td>0.94</td>
</tr>
<tr>
<td>4</td>
<td>50%</td>
<td>5%</td>
<td>0.90</td>
</tr>
<tr>
<td>6</td>
<td>50%</td>
<td>10%</td>
<td>0.80</td>
</tr>
<tr>
<td>8</td>
<td>70%</td>
<td>20%</td>
<td>0.71</td>
</tr>
<tr>
<td>10</td>
<td>40%</td>
<td>20%</td>
<td>0.50</td>
</tr>
<tr>
<td>12</td>
<td>30%</td>
<td>20%</td>
<td>0.33</td>
</tr>
</tbody>
</table>

* It is important to bear in mind that the ‘good’ outcome in each treatment is always normal health, so in the case of the first row, the calculation (under EU) is: \( 0.9 (U_{22222}) + 0.1(U_{11111}) = 0.05 (U_{dead}) + 0.95(U_{11111}) \) and assigning values of 0 and 1 to dead and normal health respectively and rearranging gives: \( 0.9 (U_{22222}) = 0.85 \), so \( U_{22222} = 0.944 \)

It should be obvious that there ought to be a systematic pattern to the series of responses given by a respondent depending on his/her valuation of state 22222 relative to dead. Specifically, if a respondent is indifferent between the two treatments in a given row, then logically they must prefer Treatment B in all the rows above this one, and prefer Treatments A in all the rows below. The row of indifference, or the row at which they ‘switch’ from column B to column A, is therefore an important source of information in the estimation of utility values. This provides a means of examining individual patterns of responses within a DCE. The methods used to model the DCE choices are outlined in section 2.4 below.

2.3 The modified SG

In the modified SG part of the questionnaire, the framing of the question was designed to closely resemble the pair-wise choices that appeared in the DCE. Rather than having the risks associated with both treatments fixed in advance and being asked to choose between treatments, only one treatment was fixed in the modified SG. Respondents were presented with a fixed risk of the health state under Treatment A, and then asked to ‘set’ that risk of death in Treatment B that made them indifferent between the two treatments. In addition, to try to make the question similar to the DCE format, a flashcard was used to graphically illustrate Treatment B for the level of risks used in the DCE (i.e. 1%, 5%, 10%, 20%, 30%, 40%, 50%, 70%, 90% risk of death). However, it was stressed to respondents that they could set the risk in Treatment B at any value, and not just one shown on the flashcard. Figure 2 shows the modified SG question used to elicit the value for health state 21121 (the flashcard is available on request from the authors). Participants were asked three modified SG questions. For health states 21121 and 22222, Treatment A involved a 90% risk of that state. For health state 22323, Treatment A involved a 20% risk of that health state, to allow for potentially lower values. The modified SG questions were asked in a fixed order 21121, 22222 and then 22323. Groups were randomized to see DCE or modified SG first.

Utility values are then estimated directly from the modified SG in exactly the same way as set out above. Considering the choice set out in Figure 2, suppose the respondent sets the indifference probability of dead at 0.20, then under EU:
0.90 (U21121) + 0.10(U11111) = 0.20 (U\text{dead}) + 0.80(U_{11111}) and assigning values of 1 and 0 to full health and dead respectively gives: (U_{21121}) = 0.78.

The format of both the modified SG and DCE questions allow worse than dead states to be valued in exactly the same manner as better than dead states.

For example, suppose the modified SG question involved a 20% risk of EQ-5D health state 22323 under Treatment A, and the respondent set the risk of death under Treatment B at 40%. Then 0.2 (U_{22323}) + 0.8(U_{11111}) = 0.40 (U_{\text{dead}}) + 0.60(U_{11111}) and assigning value of 1 and 0 to normal health and dead respectively gives: (U_{22323})=-1.

In effect, health state 22323 is worse than death, provided that the risky prospect of death is preferable to the risky prospect of health state 22323.

In the final part of the questionnaire, four questions were used to elicit participants risk attitudes for monetary lotteries, using the mid-weight method proposed by Kuilen and Wakker (2011). As this the results of the risk attitude questions are not central to the current paper, further details is available from the authors on request.

### 2.4 Modelling the DCE Choices

For this section, we introduce more formal notation. Let $X_j$ denote health-state $j$. The models developed in this section make the standard, simplifying assumption that all individuals have the same utility value for a given health state. Recall that we are commencing from the “anchors” of the utilities of normal health ($X_0$) and dead ($X_4$) being 1 and 0 respectively. There are three other health-states, $X_1$, $X_2$ and $X_3$, with utilities $u_1$, $u_2$ and $u_3$ respectively. The notation is provided in Table 2. The principal objective of the modeling is to obtain estimates of $u_1$, $u_2$ and $u_3$ i.e. to estimate utility values for health states anchored to normal health and death.

<table>
<thead>
<tr>
<th>health state</th>
<th>Definition</th>
<th>Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_0$</td>
<td>11111</td>
<td>$U(X_0)=1$</td>
</tr>
<tr>
<td>$X_1$</td>
<td>21121</td>
<td>$U(X_1)=u_1$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>22222</td>
<td>$U(X_2)=u_2$</td>
</tr>
<tr>
<td>$X_3$</td>
<td>22323</td>
<td>$U(X_3)=u_3$</td>
</tr>
<tr>
<td>$X_4$</td>
<td>Dead</td>
<td>$U(X_4)=0$</td>
</tr>
</tbody>
</table>

Consider choice problem $i$ of the DCE. The choice is between two risky treatments $A_i$ and $B_i$, defined as follows:

$A_i$: Probability $p_{a,i}$ of health state $X_{a,i}$; probability $(1 - p_{a,i})$ of health state $X_0$.

$B_i$: Probability $p_{b,i}$ of health state $X_{b,i}$; probability $(1 - p_{b,i})$ of health state $X_0$. 


Under the assumption of EU, the individual computes valuations of $A_i$ and $B_i$ as follows:

$$EU(A_i) = p_{a_i}U(X_{a_i}) + (1 - p_{a_i})U(X_0)$$
$$EU(B_i) = p_{b_i}U(X_{b_i}) + (1 - p_{b_i})U(X_0)$$

$$\Delta_i = EU(B_i) - EU(A_i)$$

(1)

Note that the symbol $\Delta_i$ is used to represent the difference in expected utilities. Let $y_i$ denote the decision. Recall that there are three possible outcomes: prefer A ($y_i = 1$); A and B equally preferable ($y_i = 2$); prefer B ($y_i = 3$). We model this decision using a version of the ordered probit model developed by Aitchison and Silvey (1957), defined as follows:

$$y_i = 1 \text{ if } \Delta_i + \varepsilon_i < \kappa$$
$$y_i = 2 \text{ if } -\kappa < \Delta_i + \varepsilon_i < \kappa$$
$$y_i = 3 \text{ if } \Delta_i + \varepsilon_i > \kappa$$

where $\varepsilon_i \sim N(0, \sigma^2)$

(2)

The parameter $\kappa$ is known as the “cut-point”, and indicates the distance from perfect indifference ($\Delta = 0$) within which “equally preferable” is reported. $\varepsilon_i$ is a normally distributed random error term.

From (2), the probabilities of the three outcomes are derived as follows:

$$P(y_i = 1) = \Phi\left(\frac{-\kappa - \Delta_i}{\sigma}\right)$$
$$P(y_i = 2) = \Phi\left(\frac{\kappa - \Delta_i}{\sigma}\right) - \Phi\left(\frac{-\kappa - \Delta_i}{\sigma}\right)$$
$$P(y_i = 3) = 1 - \Phi\left(\frac{\kappa - \Delta_i}{\sigma}\right)$$

(3)

where $\Phi(.)$ is the standard normal cumulative distribution function. From (3), the log-likelihood is constructed as follows:

$$LogL = \sum \left[ I(y_i = 1)\ln \Phi\left(\frac{-\kappa - \Delta_i}{\sigma}\right) + I(y_i = 2)\ln \left(\Phi\left(\frac{\kappa - \Delta_i}{\sigma}\right) - \Phi\left(\frac{-\kappa - \Delta_i}{\sigma}\right)\right) \right.$$

$$\left. + I(y_i = 3)\ln \left(1 - \Phi\left(\frac{\kappa - \Delta_i}{\sigma}\right)\right)\right]$$

(4)

The log-likelihood function (3) is programmed using the ML routine in STATA. The code is available from the authors on request. As normal health and death are assigned utility values of 1 and 0 respectively, the coefficients on variables $X_1$ to $X_3$ may be taken as utility values for that health state.
As mentioned previously, we also consider a non-EU theory, in the form of RDU, which allows for non-linear weighting of probabilities. Here, we assume Tversky and Kahneman’s probability weighting function (Tversky and Kahneman, 1992). If \( r \) is the probability of the good outcome (i.e. normal health in this case), then \( r \) is transformed according to:

\[
\pi(r) = \frac{r^\gamma}{[r^\gamma + (1-r)^\gamma]^\gamma}
\]

(5)

Whilst estimates of gamma are available from the literature, it is estimated within the model here. The valuations of the two treatments are derived accordingly:

\[
V(A_j) = [1 - \pi(1 - p_{a,j})]U(X_{a,j}) + \pi(1 - p_{a,j})U(X_0)
\]

\[
V(B_j) = [1 - \pi(1 - p_{b,j})]U(X_{b,j}) + \pi(1 - p_{b,j})U(X_0)
\]

\[
\Delta^{RD}_{ij} = V(B_i) - V(A_i)
\]

As in the EU model, normal health and death are assigned utility values of 1 and 0 respectively and the coefficients on the variables \( X_1 \)-\( X_3 \) may be taken to be the utility values of those states.

3. RESULTS

3.1 Estimation of utility values within the DCE model

We begin by presenting the results of the DCE models in Table 3 which shows the results under both EU and RDU. Table 3 reports the coefficients on the health state variables that represent the utility value of the associated state along with the ‘cut-point’, \( \kappa \), from the ordered probit model and the estimate of the probability – weighting factor, \( \gamma \), from the RDU model.

<table>
<thead>
<tr>
<th>Table 3: Estimates of coefficients (st errors) from DCE models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>( U_1 ) (21121)</td>
</tr>
<tr>
<td>( U_2 ) (22222)</td>
</tr>
<tr>
<td>( U_3 ) (22323)</td>
</tr>
<tr>
<td>( \sigma )</td>
</tr>
<tr>
<td>( \kappa )</td>
</tr>
<tr>
<td>( \gamma )</td>
</tr>
<tr>
<td>( N )</td>
</tr>
<tr>
<td>( \text{LogL} )</td>
</tr>
</tbody>
</table>
Note firstly that utility values derived under the assumption of RDU are considerably lower than when derived under the assumption of EU. Note also that the estimate of $\gamma$ is 0.42, an estimate that is considerably lower than that found in the literature previously. Figure 3 shows the weighting function assigned to normal health when $\gamma$ is 0.42 compared to the unweighted case represented by the diagonal line. This figure shows that the probability of normal health is seriously under-weighted when $\gamma$ is 0.42, particularly when it the probability is greater than 0.5.

**Figure 3: the probability weighting function**

3.2 Estimation of utility values from the modified SG questions
Recall that in the modified SG questions respondents were presented with a fixed risk of a health state under Treatment A, but then asked to ‘set’ that risk of death in Treatment B that made them indifferent between the two treatments. Using the EU calculations set out in the methods section above, the utility value of the health states can be calculated for each individual. Table 4 presents mean and median utility values for the 3 health states from the modified SG assuming EU preferences as is traditional.

**Table 4: Mean, median and standard deviation (SD) of utility values from modified SG assuming EU preferences**

<table>
<thead>
<tr>
<th>EQ-5D state</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>21121</td>
<td>0.909</td>
<td>0.949</td>
<td>0.118</td>
</tr>
<tr>
<td>22222</td>
<td>0.832</td>
<td>0.899</td>
<td>0.153</td>
</tr>
<tr>
<td>22323</td>
<td>0.214</td>
<td>0.500</td>
<td>0.716</td>
</tr>
</tbody>
</table>

Note that mean utility values from the modified SG are fairly close to the DCE results derived under the assumption of EU. Of course, if respondents were under-weighting probabilities and it is considered that ought to be corrected for (we will return to this in the discussion), then the modified SG results ought to ‘corrected’ also. But we do not have an estimate of $\gamma$ derived within the modified SG, so we have re-estimated the modified SG results using 1) an estimate of $\gamma$ previously found to be quite robust ($\gamma=0.65$) and 2) the value found in the RDU DCE model ($\gamma=0.42$).
Table 5: Mean, median and standard deviation (SD) of utility values from modified SG assuming RDU preferences

<table>
<thead>
<tr>
<th>State</th>
<th>Mean RDU $\gamma=0.65$</th>
<th>Median RDU $\gamma=0.65$</th>
<th>SD RDU $\gamma=0.65$</th>
<th>Mean RDU $\gamma=0.42$</th>
<th>Median RDU $\gamma=0.42$</th>
<th>SD RDU $\gamma=0.42$</th>
</tr>
</thead>
<tbody>
<tr>
<td>21121</td>
<td>0.783</td>
<td>0.803</td>
<td>0.158</td>
<td>0.541</td>
<td>0.510</td>
<td>0.201</td>
</tr>
<tr>
<td>22222</td>
<td>0.681</td>
<td>0.715</td>
<td>0.177</td>
<td>0.433</td>
<td>0.418</td>
<td>0.189</td>
</tr>
<tr>
<td>22323</td>
<td>0.528</td>
<td>0.579</td>
<td>0.232</td>
<td>0.142</td>
<td>0.138</td>
<td>0.234</td>
</tr>
</tbody>
</table>

3.4 Estimation of utility values from the 6 ‘common’ DCE questions

For the 47 respondents where it was possible to infer a value from the 6 common DCE questions involving state 22222, we calculated a mid-point value for state 22222[^4]. Recall that these 6 questions allowed a utility value (or at least a range) to be estimated at the level of the individual respondent. Under EU the mean value for state 22222 was 0.752 (sd 0.238) using this method compared to a mean value of 0.840 (sd 0.161) using the modified standard gamble. A paired t-test gave a t value of 3.224, which showed the values from the modified standard gamble were significantly higher (p value 0.002, observed mean difference 0.088 sd 0.187).

3.5 Consistency of DCE choices

We conclude by reporting the outcome of the various ‘consistency’ checks that were built into the DCE design. The most straightforward tests of consistency are the tests of dominance. Recall that all respondents were asked a different ‘risk-dominated’ question in that the same health state was involved in both treatments, but the level of risk differed. Only 2 (of 60) respondents failed this dominance test. There were also a total of 27 ‘state-dominated’ questions in which the risk level was the same but one health state was strictly better than the other (for example, 21121 is strictly better than 22222). In this context respondents were even more consistent and no respondent failed these tests of dominance. Finally, in the ‘risk/state dominated’ questions, no respondents failed these tests of dominance. Whilst this is to be welcomed, it is perhaps not too surprising that a sample of students (many of whom had studied economics) taking part in a session where two experienced moderators were on hand to answer any queries would be able to pass dominance tests in this way.

The inclusion of the 6 common questions did, however, allow a more sophisticated test of consistency in that there ought to be a given pattern of responses to these questions. Table 6 shows the distribution of responses across the 6 questions is generally as expected with the probability of respondents choosing Treatment A over B increasing as the chance of 22222 in A falls and the chance of death in B increases. As set out in the methods section, however, there should be a systematic pattern to the

[^4]: The mid-point was calculated in two ways. In method 1 we took the mid-point value between the last reported B and first reported A for everyone. In method 2, method 1 was used as before for those respondents who did not report indifference. However, for those that did report indifference we used the indifference value or mid-point of these indifference values. In both methods, for those reporting a value greater than 0.944 we took the mid-point between 1 and this value. For those reporting a value less than 0.333, we took the mid-point between this value and zero. We report here the results for the midpoint for method 1 but both gave very similar results, and both were statistically different.
series of responses made by each individual respondent depending on their evaluation of state 22222 relative to dead. Specifically, if a respondent is indifferent between the two treatments in a given row, then logically they must prefer Treatment B in all the rows above this one, and prefer Treatments A in all the rows below. Irrespective of whether a point of indifference is identified, respondents should not switch from preferring A to preferring B going down the rows.

**Table 6: Distribution of responses to the 6 ‘common’ DCE questions respondents**

<table>
<thead>
<tr>
<th>DCE Question</th>
<th>Risk of 22222 in A</th>
<th>Risk of death in B</th>
<th>Implied utility of 22222 at ‘equality’</th>
<th>Prefer A</th>
<th>Equal</th>
<th>Prefer B</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>90%</td>
<td>5%</td>
<td>0.94</td>
<td>20.0%</td>
<td>11.7%</td>
<td>68.3%</td>
</tr>
<tr>
<td>4</td>
<td>50%</td>
<td>5%</td>
<td>0.90</td>
<td>38.3%</td>
<td>8.3%</td>
<td>53.3%</td>
</tr>
<tr>
<td>6</td>
<td>50%</td>
<td>10%</td>
<td>0.80</td>
<td>41.7%</td>
<td>25.0%</td>
<td>33.3%</td>
</tr>
<tr>
<td>8</td>
<td>70%</td>
<td>20%</td>
<td>0.71</td>
<td>48.3%</td>
<td>25.0%</td>
<td>26.7%</td>
</tr>
<tr>
<td>10</td>
<td>40%</td>
<td>20%</td>
<td>0.50</td>
<td>71.7%</td>
<td>15.0%</td>
<td>13.3%</td>
</tr>
<tr>
<td>12</td>
<td>30%</td>
<td>20%</td>
<td>0.33</td>
<td>86.7%</td>
<td>6.7%</td>
<td>6.7%</td>
</tr>
</tbody>
</table>

When the individual pattern of choices is examined in more detail, however, a number of logical consistencies arise in the series of responses given. We have classified respondents’ series of answers into 3 ‘types’ which we explain below.

- Type 1 move from preferring B to A down the series of questions (may or may not have an ‘equals’ response where ‘switch’ made).
- Type 2 select ‘equally preferred’ for more than one question.
- Type 3 move from preferring A to B moving down the rows- or move from preferring B to A then back to B again down the rows.

Whilst Type 1 respondents are consistent, Types 2 and 3 could both be thought of as inconsistent in their pattern of responses although the degree of inconsistency is different between the two types. Type 2 respondents may simply be demonstrating that they can only identify a range within which their true value of 22222 lies and hence have chosen ‘equally preferred’ more than once. Type 3 respondents are strictly inconsistent in their pattern of responses and it is impossible to determine even a range within which their value of 22222 lies. Of the 60 respondents 38 (63%) were Type 1, 9 (15%) were Type 2 and 13(22%) were Type 3. Hence, whilst only 22% were strictly inconsistent in their series of choices, only 63% were strictly consistent. This compares with almost all respondents passing the more straightforward dominance tests. If one argument for the use of DCEs over methods such as SG (or TTO) is that respondents have found to be inconsistent when completing tasks such as SG, it is obviously important that responses to DCE questions are subjected to the same level of scrutiny.
6. DISCUSSION

We report the results of an exploratory study that set out to develop a risk-based DCE approach to derive utility values for health states. Health states are anchored to normal health and death allowing utility values to be derived directly within the DCE. The design allows worse than dead states to be valued in the same manner as better than dead states. Whilst the nature of the risk attribute used in a previous risk-based DCE study was such that direct comparisons with SG were problematic (van Houtven et al, 2011), we set out here to make the methods as comparable as possible.

Our results show a broad correspondence between the results from DCE model and the mean (modified) SG results, particularly under the assumption of EU preferences. The results are very similar indeed for two of the health states (21121 and 22222) whilst the DCE model results are higher than modified SG for 22323. It would be interesting to see what pattern would emerge should a wider range of health states be evaluated.

Adjusting both the DCE and modified SG results for probability weighting significantly reduces utility values. Indeed we found very pronounced under-weighting of the ‘better’ outcome in our study and the downward adjustment of utilities under RDU resulted in values that many readers may find implausible. This seems particularly likely for state 21121 which has a utility value from the RDU DCE model of 0.541. A value that suggests a state that involves some problems walking about and moderate pain and discomfort is almost half-way to being dead just seems implausibly low. The marked difference between the values derived under RDU and EU would seem to rule out that the former are low simply because they were derived in a student sample.

It is probably worth noting, however, that the apparent marked under-weighting of the ‘better’ outcome here may have been a framing effect specific to the study design here. The ‘better’ outcome here was always normal health and it seems plausible that information on the probability of that outcome may have essentially been ‘edited out’ as a simplifying heuristic deployed by respondents. This is obviously something that can be tested in future studies by allowing the ‘better’ outcome to vary across treatments.

Irrespective of the actual values derived here, the more general issue is whether utility values ought to be adjusted for probability weighting (or indeed loss aversion) if that is found to exist. Clearly this is a normative judgment that cannot be made post-hoc after estimates are derived. It has been argued that, even if found not to be descriptively valid, EU should still be used as the normative basis on which to evaluate policies (de Palma et al, 2008). The main justification for this is that, “when given enough opportunity to learn about the consequences of non-EU decision making, most people switch to EU behavior.” P 283. But this does not rule out the use of non-expected utility models in order to adjust responses in order to ‘correct’ for factors such as probability weighting and loss aversion. Indeed it is widely considered that SG values ought to be corrected for probability-weighting if utilities are not to be

We purposely resist comparing the values derived here to the EQ-5D ‘tariff’ values based on TTO as that is to imply a gold standard that is not appropriate.
biased upwards (see, for example, Doctor et al, 2010). Further discussion of these normative issues is beyond the scope of the current paper, but it is clearly an important consideration whenever the choice of model has a significant impact on estimated values.

It is clear from our results, however, that the issue of the appropriate model of risky choice applies equally whether a modified SG or risk-based DCE approach is taken. Hence, criticisms of the SG on the grounds EU is not descriptively valid does not, in itself, provide a rationale for the use of DCE. Utility values derived via SG may be adjusted for probability weighting and loss aversion if that is considered appropriate. The difference between the RDU DCE model estimates and our SG results assuming RDU was that in the latter the probability weighting factor had to be derived elsewhere. Although not a prominent part of this paper, we did explore the use of risk attitude questions that may allow a within-sample probability weighting factor to be derived and used to adjust SG valuations. An obvious methodological issue there would be whether risk attitudes in the domain of money lotteries would necessarily be the same as those in health (see Prosser and Wittenberg, 2007) and the feasibility of being able to ask the required number of such questions in a population sample. This remains an area for future research although authors have had some success in this area (e.g. Booij et al, 2010).

There is clearly also scope for further development and refinement of the econometric model demonstrated here. For example, we made here the simplifying assumption that all respondents have the same utility value for a given health state, and that, in the RDU model, all respondents have the same probability weighting parameter. The fact that repeated decisions are made by each respondent enables the estimation of heterogeneity parameters representing between-respondent variation in utility and/or probability weighting parameters. This extension is an interesting possibility for future research.

There are, of course, other models of risky choice that may be applied to the data, such as those allowing for loss aversion raising the issue of what the appropriate reference point is. It could be that the outcome common to all choices, in this case normal health will be taken as the reference point and, hence, all moves considered as losses. Alternatively, it could be that respondent’s own health would be taken as the reference point and gains and losses assessed relative to that. Given the demographic of our sample, any deviation between ‘own’ and ‘normal’ health is unlikely to be significant. Importantly for the analysis conducted here, if the questions were seen either entirely in the loss or gain domain, adding loss aversion into our model is unlikely to affect our results. But identification of the reference point, and the need to model the treatments as gains and losses is likely to be an important issue in a population sample.

Likewise, TTO values may easily be adjusted for time preference, but that would require applying a discount rate taken from elsewhere or the inclusion of additional discounting questions alongside the TTO exercise.

---

6 Likewise, TTO values may easily be adjusted for time preference, but that would require applying a discount rate taken from elsewhere or the inclusion of additional discounting questions alongside the TTO exercise.
An obvious limitation of the study is that we used a convenience sample of students as it was largely exploratory. Whilst the use of interactive computerized programmes would make the tasks more ‘user-friendly’ than the paper-based questionnaire deployed here, it remains an open question how members of the public would cope with the exercise. This is an area for future research. Our results do show that, even in a sample that could reasonably be expected to be more consistent than a general population one, closer scrutiny of the DCE choices shows that only two thirds of respondents had a pattern of responses that was strictly consistent. Whilst we do not conclude anything definitive from this finding, we do believe that it is important to explore individual patterns of responses within DCEs. This is particularly so if it seems that errors made are not in fact ‘random’ at all, but, rather, reflect the same patterns of anomalies and cognitive biases that are uncovered in value elicitation studies elsewhere.

We conclude by making some more general points about the assessment of the relative merits of DCE and more traditional methods such as SG and TTO. Methods such as SG and TTO are traditionally thought of as ‘matching’ techniques— whereby the task is to ‘set’ the level of risk/duration that makes the respondent indifferent between two options. There is a literature on the fact that ‘matching’ and ‘choice’ tasks maybe tapping into different cognitive processes and, hence, the results are likely to differ across methods. One criticism of ‘matching’ tasks (e.g. Tversky et al, 1988), is that, asking respondents to ‘match’ on any single dimension encourages respondents to attach undue weight to that specific dimension while neglecting other factors that they would otherwise wish to be taken into consideration.

Whilst the modified SG that respondents completed here was an actual ‘matching’ task (in that we asked respondents to directly ‘set the probability of death in Treatment B to make them indifferent between A and B), it is important to acknowledge that most SG and TTO elicitation techniques actually present respondents with a series of pair-wise choices. Holding the format of the questions the same, the only difference between SG (or TTO) and a DCE is that in the former the choices are generally generated by an interactive process that tries to ‘hone in’ on that respondents point of indifference. As the purpose is to find that point where the decision is the most difficult, it is somewhat spurious to criticize the SG (and TTO) on the basis that respondents find them more difficult than DCE (see, for example, Stolk et al, 2010 and Ratcliffe et al, 2011). When considered in this way, the SG (and TTO) could be seen as more ‘efficient’ techniques at arriving at utility values than DCE. On the other hand, a possible drawback of any interactive approach is that ‘starting point’ biases and anchoring effects may be introduced that are avoided in the DCE.

We believe the body of work currently being undertaken to estimate utility values within a DCE is important and that the method does have the potential to replace traditional methods such as SG and TTO. A number of methodological issues remain in deriving utility values within a DCE, however, and the likely advantages of the approach over more traditional methods needs to be better set out.

---

7 This literature is too vast to cover here, but a good summary of cognitive biases and anomalies is given in Loomes, 2006.

8 We realise, of course, that DCEs are generally setting out to look at a wider range of attributes than are typically included in SG and TTO studies—this comment refers to situations in which a DCE would be used to try and replicate SG and/or TTO.
**Acknowledgements**

We are grateful to the 60 students who took part in the experiment. We would also like to thank Tim Miller from the Finance and Economics Experimental Laboratory at Exeter (FEELE) based at the University of Exeter, who helped recruitment in that location and Todd Kaplan who is the academic lead for the laboratory. We are grateful to Jose-Luis Pinto-Prades who carried out supplementary analysis of the data collected in order to look at the pattern of choices more carefully at the individual level. We thank members of the Health Economics Group at Exeter, and for those attending seminars in Sheffield and Bristol Universities, for their constructive comments and advice. We are also grateful to the discussant and participants at the Health Economics Study Group in Oxford for their valuable comments. The study was funded by internal funds of the School of Economics and Finance at Queen Mary University. There are no conflicts of interest.

**References**


Brazier J, Rowan D., Yang, Tsuchiya A. Comparisons of health state utility values derived using time trade-off, rank and discrete choice data anchored on the full health-dead scale. *Eur. J. Health Econ*. Forthcoming


