HEALTH ECONOMICS GROUP
Faculty of Medicine and Health Norwich
Medical School

University of East Anglia

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

Angela Robinson[1], Anne Spencer[2] & Peter Moffatt[1]

[1.] University of East Anglia
[2.] Peninsula Medical School, University of Exeter.

July 2012 HEG working paper 12-01

JEL Classification: I10, D81

Keywords: Discrete choice experiments (DCE), Risky choices, Health state utility values, Expected utility theory, Rank dependent utility theory

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. The study was funded by internal funds of the School of Economics and Finance at Queen Mary University and there are no conflicts of interest.

Contact information: Angela Robinson, Health Economics Group, Norwich Medical School, UEA, Norwich, NR4 7TJ, E-mail: Angela.Robinson@uea.ac.uk

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

Angela Robinson[1], Anne Spencer[2] & Peter Moffatt[1]

July 2012

ABSTRACT

Discrete choice experiments (DCEs) allow a number of characteristics to be traded-off against one another. An overriding methodological challenge faced is how best to apply DCEs to questions involving those attributes commonly used in value elicitation exercises such as risk, time (Bansback et al. 2012) and numbers treated (Robinson et al, 2010). Flynn (2010) concluded that in developing the methods, it was important to understand more fully the preferences of individual respondents. The study reported here sets out to provide such insights by enhancing a DCE design with additional questions that allow utility values to be derived at the individual level also.

The DCE presented respondents with eight pairwise risky choices to estimate aggregate utility values for three EQ-5D health states, ranging from mild to severe. The design allowed the elicitation of utility values for worse-than-dead states. Risk was represented using the stimulus used by EuroVaQ (http://research.ncl.ac.uk/eurovaq/). Three main devices were used to collect additional individual level data. Firstly we included six additional DCE questions that were not used to estimate the aggregate DCE model but allowed the utility value of one health state to be estimated at the level of the individual. These six questions provided more extensive tests of the internal consistency of the pairwise choices undertaken in the DCE. Secondly, respondents were asked three questions where the risk in one of the two treatments was fixed, and they set the risk of the other treatment (a modified SG question). These questions then allowed us to estimate utility values for all three health states. Finally, we collected respondents risk attitudes using Kuilen and Wakker's 2011 measure.

We collected data on a convenient sample of 59 students studying Economics or Geography at the University of London and Exeter in 2011/12.

Preliminary results show that 22 of the 59 respondents gave a series of DCE responses that were internally inconsistent. We report here the implications of the results for the inclusion of risk as an attribute in DCEs and for preference elicitation more broadly.

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 those involving the services received[1,2]. In many such applications, these attributes can reasonably be assumed to contribute to utility in an additive way, albeit with the possibility of interaction effects between attributes. There has, however, been recent interest in using the DCE method to derive health state utilities thereby requiring the inclusion of attributes commonly used in value elicitation exercises such as risk, duration[3] and numbers treated[4]. 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[5].

Another challenge in using DCEs to derive health state utilities is in anchoring values to normal health and dead. Whilst the possible advantages of DCE over traditional methods – such as standard gamble (SG) and time trade off (TTO) have been set out previously[6], relatively little is known about how the technique performs in head-to- head comparisons with SG or TTO. Bansback et al have previously used a DCE to elicit utility values using ‘TTO style’ questions in order to value a range of EQ-5D states[3]. To our knowledge no previous researcher has set out to use a DCE to elicit utility values using risky ‘SG-like’ questions.

The study reported here investigates the use of risk-based DCEs to elicit health state utility values and adds to the small existing literature on the use of DCEs in this context.

2. METHODOLOGICAL ISSUES TO BE ADDRESSED

Estimating utility values for health states directly from a DCE model requires health states 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[7] but the results cannot then be used as utility values and incorporated into QALY calculations. We were keen to include both normal health and dead directly in the model, but the inclusion of the state ‘dead’ in a DCE is potentially problematic. Previous DCE studies that have included ‘immediate death’ as a state (and so exclude survival as an attribute) have been criticized by Flynn et al on the basis that they do not allow for trade-offs between quality of life and length of life[8]. When survival is not an attribute most people may choose life over ‘immediate death’, making it hard to estimate states close to dead. In addition, people are likely to make fewer errors when their preferences are well defined, as they will be for mild or severe states, and make more errors for the states in-between. Such variation in errors will violate the assumption of constant error variance typically applied in DCEs.

As well as allowing health states to be ‘anchored’, the inclusion of dead as an attribute allows states worse-than-dead to be valued on the same scale as better than dead states. Traditional procedures for valuing states worse than dead represent a fundamental departure from those used to value better than dead states and there is a large body of evidence that shows responses can be affected by descriptive and procedural invariance[9] and we argued previously[10] 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 measured states worse-than-dead and better than dead using a common elicitation procedure now commonly referred to as the lead time approach[10].

Previous DCE studies have linked health states to normal health and dead by either including ‘immediate death’ as a state or survival duration as an attribute. In the first approach the survival duration in each state can be fixed and the values of health states are rescaled to normal health and dead by dividing estimated coefficients for states by the coefficient on the ‘dead’ state[6]. The other approach, termed DCE_{TTO}, includes health states and survival duration (not including zero) as attributes, but no longer includes ‘immediate death’[3]. In the DCE_{TTO}, normal health is set at one and values worse than dead can be inferred ‘indirectly’ at a sample level, by looking at reductions in quality of life, from levels from 1 to 2 and 2 to 3 between states. For example, setting normal health equal to 1, and subtracting the impact of decreases in quality of life as you move through the levels, there will come a point that when the values lie below. It could be argued, however, that direct comparisons against immediate death are important to check the validity of worse than dead scores.

Whilst there are inherent difficulties in including the state ‘dead’ in a study that also includes duration as an attribute, there is no obvious reason why there would be similar difficulties in using risk as an attribute. An approach using risk, rather than duration lends itself more readily to the inclusion of ‘immediate death’ as an attribute. Comparing two risky treatments, links to the work by Keeney and Raifa[11] and has been used successfully in other studies[12]. There is a large and growing body of research within economics looking at decision making under risk[13,14]. Moreover, the potential biases involved in using risky choices are well documented and there has been relative success in adjusting for these biases in risky choices[15,16].

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 that McFadden and Heckman developed 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. Cumulative Prospect Theory (CPT) assumes a probability weighting function and also allows for people to experience greater changes in utility from losses compared to gains (loss aversion).

An important step forward has been made by de Palma et al who argued that these non-expected utility functions can also be used to underpin random utility theory[17]. They outline the types of data needed to infer these models. However, like others they argue that expected utility should still be used as the normative basis on which to evaluate policies. 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[17]. . These non-expected utility models therefore are simply used to adjust responses to take account of probability weighting and loss aversion so that they can be used to evaluate competing policies.

One previous 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. The DCE included three life-threatening side effects, which were reported as 10-year mortality risk of 0.5%, 2% and 5%[18]. They found evidence of non-linearity in how these risks were perceived, and that this non-linearity varies across side effects. When they applied a Rank Dependent Utility model to the DCE data, which allowed for probability weighting, they found lower utility values than when EU was assumed.

This led them to argue that traditional SG methods assuming linear risk preferences are biased but made the point that it was difficult to do a direct comparison of utility values 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 DCE to look at people's willingness to accept the risks involved in SG questions that are typically used to elicit the values of different health states.

Of course, one feature of the DCE approach that distinguishes it from methods such as SG and TTO is that it is inherently an aggregate method. Whilst SG and TTO allow values to be elicited at the level of the individual respondent by eliciting individual points of indifference, no individual point of indifference is generally attained in DCEs. This fundamental difference makes comparison between methods difficult.

The aims of this research are therefore:

1) To elicit 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.

3. METHODS

3.1 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 participants were told that the Government and other bodies wanted some guidance from members of the public about priorities for funding different treatments. 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. 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 and made use of money lotteries. All methods are explained in detail below.

3.2 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 dead) and an associated chance, 1-p, of normal health (11111). Thus, normal health appeared in all treatments and was coloured pink. A typical question is shown in Figure 1 and used graphical displays to illustrate risk information (developed by EuroVaQ http://research.ncl.ac.uk/eurovaq/). 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 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 state which treatment they preferred by ticking 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[19-22]. It is important to stress here that we are *not* setting out to derive weights for the

EQ-5D descriptive system as setting out to do so would obviously require each dimension- and level- to be modeled. We are simply setting out to value 3 health states here and, hence, our design is a very simple one involving only two attributes- health state and risk. The DCE questions varied on one or more of the two attributes shown below:

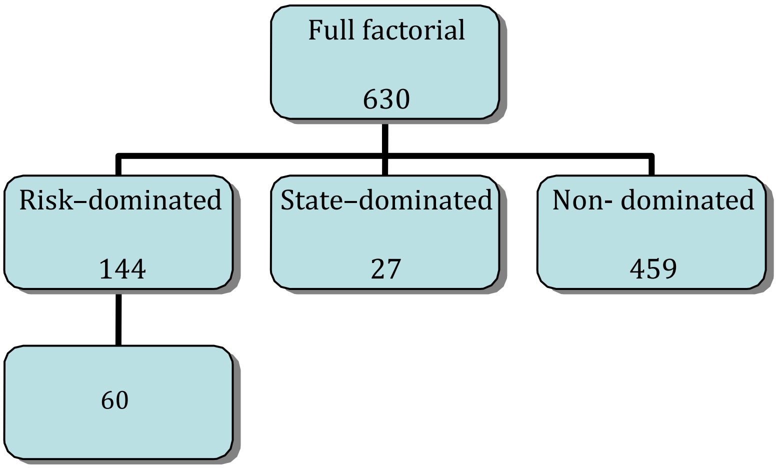
- The health state (either 21121, 22222, 22323 or dead) coloured yellow, green, taupe and blue respectively.
- The probability of that outcome (either 1%, 5%, 10%, 20%, 30%, 40%, 50%, 70%, 90%).

The attributes and levels set in this study produced a total of 630 different combinations[1]. Given the uncertainty surrounding use of optimal design for multiplicative model[5] we chose to include all non-dominated combinations. Of the 630 combinations, 144 combinations were '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. Such questions do not provide any information about the nature of trade-offs between health and risk but can be useful as 'consistency' checks, so we elected to ask all participants one of the 144 'risk-dominated' questions. As each participant was presented with a different 'risk-dominated' question, 60 of the 144 questions of this type were therefore included in the design.

As there is a 'logical' ordering of health states in that $21121 \succ 22222 \succ 22323$, there is another type of dominance that we term 'state-dominance'. 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 comparisons. These 27 questions were randomly allocated across the 60 respondents along with the remaining 459 non–dominated choices. Thus, a

[1] The total number of scenarios was $9\times1\times4=36$. The number of ways of choosing r=2 scenarios at random from n=36 is $(n)!/(n-r)!r!=36.35/2=630$.

total of 546 (60+27+459) of the 630 choices in the full factorial were used in the design.



The first question in the DCE part of the questionnaire was one drawn randomly from the 144 'risk dominated' comparisons described above. Questions 3, 5, 7, 9, 11 and 13–15 were eight drawn randomly from the remaining 486 questions from the full factorial design. Questions 2, 4, 6, 8, 10 & 12 were six questions that were also from the full factorial but in this case the same set of 6 was presented to all respondents. 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. 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).

Table 1: The 6 DCE questions answered by all respondents

Question	Risk of 22222* in Treatment A	Risk of death* in Treatment B	Implied value of 22222 under EU
2	90%	5%	0.94
4	50%	5%	0.90
6	50%	10%	0.80
8	70%	20%	0.71
10	40%	20%	0.50
12	30%	20%	0.33

* 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 depending on the respondent's 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. The methods used to model the DCE data are included along with those results in section 5 below.

3.3 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, in modified SG respondents were presented with a fixed risk of the 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. We use the term ‘modified’ SG to denote that, unlike conventional SGs that generally involve certainty, risk was associated with both treatments here. As above, this was done in order to avoid ‘certainty bias’ but also to make the questions resemble as closely as possible those used in the DCE questionnaire. Figure 2 shows the modified SG question used to elicit the value for health state 21121. 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 :

$0.90\ (U_{21121}) + 0.10(U_{11111}) = 0.20\ (U_{dead}) + 0.80(U_{11111})$ and assigning values of 1 and 0 to full health and dead respectively gives: $(U_{21121}) = 0.78$.

Modified SG values were also estimated assuming RDU. In RDU people overweight small probabilities and underweight large probabilities. Although there are many versions of this weighting function that can be imposed, most research has found the weighting function developed by Tversky and

Kahneman[23] to be quite robust. Their weighting function □(p) is shown in equation (5) below and we initially assume a value of ©=0.65 from the literature[23]. As the name RDU suggests, outcomes are first ranked from best to worse, before applying the weighting function. In the simple case of just two outcomes, where normal health is always the better outcome, normal health is given a weight of □(p) and the other outcomes a weight of 1-□(p).

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, which we have argued previously is a desirable feature of a utility elicitation technique[10]. Suppose in the modified SG that the respondent set the indifference risk of death in Treatment B greater than the risk of the health state in Treatment A, this is to value the heath state in A as worse than dead. Suppose in the modified SG question involving a 20% risk of EQ-5D health state 22323 under Treatment A, that the respondent set the risk of death under Treatment B at 40%. Then 0.2 (U_{22323}) + 0.8(U_{11111}) = 0.40 (U_{dead}) + 0.60(U_{11111}) and assigning value of 1 and 0 to normal health and dead respectively gives: (U_{22323})= -1

Often overlooked is the fact that the lower bound of worse than dead scores are always affected by the stimulus presented to respondents and that is no different here. The lower bound of the modified SG scores was determined by the risk of the health state under Treatment A (90% for 21121 and 22222 and 20% for 22323). It is easy to work out that the lower bound for health state 22323 is -4 whilst for states 21121 and 22222 it is – 0.11. The DCE design has a lower limit of -89, which was the maximum lower bound given the choice of levels on the risk attribute[2].

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[24]. As the results of the risk attitude questions are not central to the current paper, we present these questions in the appendix.

4. RESULTS

4.1 The modified SG questions

We begin with the results of 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 2 presents mean and median utility values for the 3 health states from the modified SG assuming both EU and RDU preferences (weighting function = 0.65).

[2] The DCE design outlined on page 5 shows that the at the extreme a 90% chance of dead on one side could be involved in a pair-wise comparison with a 1% chance of, for example 22323 on the other, in which case, at indifference; 0.01 (U_{22323}) + 0.99(U_{11111}) = 0.90 (U_{dead}) + 0.10(U_{11111}) and assigning values of 1 and 0 to normal health and dead respectively: (U_{22323})= -89. Hence, -89 is the natural lower bound on worse than dead values in the DCE design.

Table 2: Mean, median and standard deviation (SD) of utility values from modified SG assuming EU and RDU(©=0.65)

EQ -5D state	Mean EU	Median EU	SD EU	Mean RDU ©=0.65	Median RDU ©=0.65	SD RDU ©=0.65
21121	0.909	0.949	0.118	0.783	0.803	0.158
22222	0.832	0.899	0.153	0.681	0.715	0.177
22323	0.214	0.500	0.716	0.528	0.579	0.232

4.2 Consistency of DCE responses

We turn now to the results of the DCE questions and begin by reporting the results of the consistency tests that were built into the design (the modeling results are presented later). 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 this dominance test. 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. Table 3 presents the pattern of responses to these 6 questions.

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 above, there should be a systematic pattern to the 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.

Table 3: Distribution of responses to the 6 'common' DCE questions respondents

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

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 B to A and then back to B (or vice versa) down the series of questions.

Whilst Type 1 are consistent respondents, 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.

Where it was possible to infer a value from the 6 DCE questions (i.e. type 1 and 2) we calculated a mid point value for state 22222[3]. We then compared this value against the value inferred from the modified standard gamble. The mean difference between the value inferred from the 6 questions and the modified standard gamble was 0.088 (sd 0.187). A paired t-test gave a t value of 3.224, which was not statistically different (p value 0.998).

[3] The midpoint was calculated in two ways. In method 1 we took the midpoint 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 midpoint of these indifference values. In both methods, for those reporting a value greater than 0.944 we took the midpoint between 1 and this value. For those reporting a value less than 0.333, we took the midpoint 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 not statistically different.

5. MODELLING THE DCE CHOICES

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. Definitions of the five health states are provided in Table 4. The principal objective of the modeling is to obtain estimates of u_1, u_2 and u_3.

Table 4: Notation used in the DCE model

health state	***Definition***	***Utility***
X_0	11111	$U(X_0)=1$
X_1	21121	$U(X_1)=u_1$
X_2	22222	$U(X_2)=u_2$
X_3	22323	$U(X_3)=u_3$
X_4	Dead	$U(X_4)=0$

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:

$$\begin{aligned} EU(A_i) &= p_{a,i}U(X_{a,i}) + (1-p_{a,i})U(X_0) \\ EU(B_i) &= p_{b,i}U(X_{b,i}) + (1-p_{b,i})U(X_0) \qquad (1) \\ \Delta_i &= EU(B_i) - EU(A_i) \end{aligned}$$

Note that the symbol $\Box_i$ 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[25], defined as follows:

$$y_i = 1\ if\ \Box_i + \sum_i < |$$
$$y_i = 2\ if\ \Box\!| < \Box_i + \sum_i < |$$
$$y_i = 3\ if\ \Box_i + \sum_i > |$$

where $\sum_i \sim N\left(0, ſ^2\right)$ (2)

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

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

$$P\left(y_i = 1\right) = \surd\left(\frac{\Box\!|\ \Box\Box_i}{ſ}\right)$$
$$P\left(y_i = 2\right) = \surd\left(\frac{|\ \Box\Box_i}{ſ}\right)\Box\surd\left(\frac{\Box\!|\ \Box\Box_i}{ſ}\right) \qquad (3)$$
$$P\left(y_i = 3\right) = 1\Box\surd\left(\frac{|\ \Box\Box_i}{ſ}\right)$$

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

$$LogL = \underset{i}{\Box}\left[\begin{array}{l} I\left(y_i = 1\right)\ln\surd\left(\frac{\Box\!|\ \Box\Box_i}{ſ}\right) + I\left(y_i = 2\right)\ln\left[\surd\left(\frac{|\ \Box\Box_i}{ſ}\right)\Box\surd\left(\frac{\Box\!|\ \Box\Box_i}{ſ}\right)\right] \\ +I\left(y_i = 3\right)\ln\left[1\Box\surd\left(\frac{|\ \Box\Box_i}{ſ}\right)\right] \end{array}\right] \qquad (4)$$

The log-likelihood function (3) is programmed using the ML routine in STATA. The code is available from the authors on request.

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[23] probability weighting function. If p is the probability of the good outcome (i.e. normal health), then p is transformed according to:

$$\square(p) = \frac{p^{©}}{\left[p^{©} + (1 \square p)^{©}\right]^{1/©}} \qquad (5)$$

The valuations of the two treatments are derived accordingly:

$$V(A_i) = \left[1 \square \square\left(1 \square p_{a,i}\right)\right] U\left(X_{a,i}\right) + \square\left(1 \square p_{a,i}\right) U\left(X_0\right)$$

$$V(B_i) = \left[1 \square \square\left(1 \square p_{b,i}\right)\right] U\left(X_{b,i}\right) + \square\left(1 \square p_{b,i}\right) U\left(X_0\right)$$

$$\square_i^{RD} = V(B_i) \square V(A_i)$$

Results from both EU and RDU models are shown in Table 5.

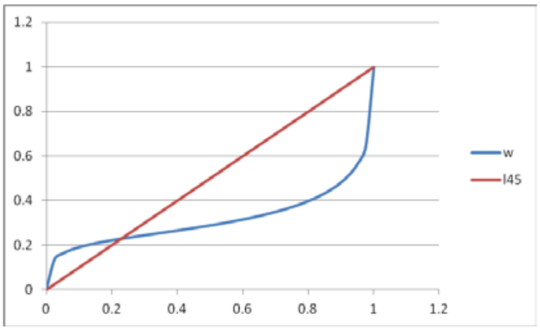
Table 5: Estimates of coefficients (st errors) from DCE models

	EU	RD
U_1 (21121)	0.907(0.029)	0.510(0.054)
U_2 (22222)	0.789(0.015)	0.311(0.040)
U_3 (22323)	0.284(0.050)	0.090(0.027)
ʃ	0.154(0.010)	0.132(0.010)
|	0.031(0.003)	0.031(0.003)
©		0.419(0.029)
N	900	900
LogL	-613.53	-544.31

Note firstly that the estimate of © is 0.42. This implies the probability weighting function shown in Figure 3. Note also that this function represents the weighting function assigned to the probability of the best outcome (i.e. normal

health), this being measured on the horizontal axis. It is seen that the probability of normal health is seriously under-weighted, particularly when it is greater than 0.5. In the simple case of just two outcomes, where normal health is always the better outcome, normal health is given a weight of □(p) and the other outcomes a weight of 1-□(p).

Figure 3: the probability weighting function



Note secondly that estimates of utilities are considerably lower under RDU than under EU. This is because EU disregards the serious under-weighting of normal health that is evident in Figure 3, and therefore seriously under-estimates the utility of this health state, relative to the other health states. It is probably worth mentioning, however, that the results under EU for 21121 and 22222 in particular may seem the more plausible to readers.

We turn now to a comparison of the results of the DCE with the modified SG. As above, the modified SG data were analysed under both EU and RDU theory assuming a value of ©=0.65 taken from the literature. Following the results from our RDU DCE model given above, we also estimated the modified SG data using ©=0.42. The modified SG results are given in Table 6 whilst Table 7 compares the aggregate results from the DCE models with those of the modified SG.

Table 6: Modified SG mean, median, standard deviation (SD) under EU and RDU

State	Mean EU	Median EU	SD EU	Mean RDU ©=0.42	Median RDU ©=0.4	SD RDU ©=0.42	Mean* RDU ©=0.65
21121	0.909	0.949	0.118	0.541	0.510	0.201	0.783
22222	0.832	0.899	0.153	0.433	0.418	0.189	0.681
22323	0.214	0.500	0.716	0.142	0.138	0.234	0.528

*The medians and standard deviations are already reported in Table 2 above.

Table 7: Comparison of DCE and Modified SG means

State	DCE EU	DCE RDU (©=0.42)	SG EU	SG RDU (©=0.42	Difference EU	Difference RDU (©=0.42
21121	0.907	0.510	0.909	0.541	-0.002	-0.031
22222	0.789	0.311	0.832	0.433	-0.043	-0.122
22323	0.284	0.090	0.214	0.142	0.070	-0.052
				mean	0.008	-0.068
				SD	0.057	0.048
				Pearson correlation	0.997	0.974

It can be seen that under EU the difference between the results of the DCE model and the modified SG are reasonably close to one another. The mean absolute difference between DCE and modified SG under EU is quite small 0.008 (SD 0.057) and positive. The mean absolute difference between DCE and modified SG under RDU (©=0.42) is larger (0.068) and negative. It is important to remember, however, that our estimate of ©=0.42 came from the

DCE model itself- as Table 6 shows- had we used an estimate from the literature of ©=0.65 the results under RDU would be very different. This highlights the importance of applying the same model of risky choice before comparing across methods.

6. DISCUSSION

We report the results of an exploratory study that set out to use a risk-based DCE to assess utility values for 3 EQ-5D health states. The design allows states to be anchored to normal health and death allowing utility values 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. Whilst the nature of the risk attribute used in a previous risk-based DCE study was such that comparisons with SG were problematic, we strove here to make the methods as comparable as possible. Hence, we believe our study offers the first opportunity to make a meaningful comparison of DCE results with those from (modified) SG. Another strength of our study is the inclusion of the 6 DCE questions that allowed an examination of individual patterns of responses to risk-based DCE questions.

Our results show a broad correspondence between the results from DCE 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 results are higher than SG for 22323. It would be interesting to see what pattern would emerge were a wider range of health states evaluated. A greater number of health states were examined by

Stolk[26] and Brazier et al (forthcoming)[6] who found that the results of DCE based on TTO resulted in higher valuations than TTO questions.

It is clear from our results that assumptions made about risky choice behaviour is very important and that applies whether a modified SG or DCE approach is taken. There are, of course, other such models that may be applied to the data and we intend to explore this in future. Although not a prominent part of this paper, we have the data from our risk attitude questions discussed in the appendix and we would ideally want to try and see to what extent that data could inform the choice of models. 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[27]. We welcome discussion from HESG members on the application of other models non EU models of risky choice.

In the econometric model developed in section 5, we made 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.

We return to the comparison of the DCE and modified SG results. Of course, observing differences (or similarities) in results across methods does not, in itself, tell us anything about which set of results is ‘better’ than the other. As DCE is inherently an aggregate modeling approach whilst SG (and TTO) sets out to estimate points of indifference at the level of the individual before taking some measure of central tendency, comparing approaches is obviously problematic. As above, the model of risky choice is equally important in both SG and DCE, so the assessment clearly cannot be made on that basis.

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[9]), 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. When considered in this way, it could be argued that SG (and TTO) are more 'efficient' techniques at arriving at utility values[4].

And using an actual 'matching' process as we did here in our modified SG, we derived utility values for the 3 health states using only 3 questions.

This brings us to the issue of study design. We acknowledge that the design applied here was very simplistic in that it set out to use the full factorial minus a number of dominated options. Although many 'dominated' choices were removed, there were still a large number of choices that were essentially 'redundant' in that they were asking for preferences over treatments that were almost certainly very far apart indeed in terms of utility space. For example, depending on the level on the risk attributes appearing under Treatments A and B, a number of choices were essentially asking whether state 21121 was worse than dead. We believe that most readers would not consider it a good use of resources to present respondents with many such questions and indeed no respondent gave a response that indicated they believed 21121 to be worse than dead[5]. We need to stress that we are not making a criticism of the DCE approach in general here- we appreciate this is an issue of our study design and that more sophisticated methods may have been used. But the full factorial in studies such as ours will always generate many choices that are likely to be very 'far apart' in utility space and that is an important point to make in considering the use of DCEs in utility assessment.

[4] 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.

[5] We are grateful to Jose-Luis Pinto-Prades who carried out some supplementary analysis of our data set. Although not included here, his analysis showed that many of our questions yielded little information and we intend to look into these issues further.

The example we give above of questions comparing 21121 with death may be linked to the point made previously by Flynn et al[8] about including death in particular in a DCE, and that remains an important methodological point, but the issue is obviously a more general one. Depending on the level on the risk attributes appearing under Treatments A and B, a number of choices were also asking whether state 21121 was worse than 22323. Again, we believe that most readers would not consider asking many such questions a good use of resources. It was evident from watching respondents complete their booklets that many choices were 'very easy' indeed. As above, Flynn et al [8] argue that any variation in errors will violate the assumption of constant error variance typically applied in DCEs. Whilst further discussion of study design issues are beyond the scope of this paper, it seems likely that designs based on an 'utility balance' approach may be the way forward. We believe the results of this exploratory study may be used in generating such a design for future use.

References

1. Damman OC, Spreeuwenberg P, Rademakers J, Hendriks M. Creating Compact Comparative Health Care Information: What Are the Key Quality Attributes to Present for Cataract and Total Hip or Knee Replacement Surgery? *Med. Decis. Mak.* Mar-Apr 2012;32(2):287-300.

2. Scott A, Watson MS, Ross S. Eliciting preferences of the community for out of hours care provided by general practitioners: a stated preference discrete choice experiment. *Soc. Sci. Med.* Feb 2003;56(4):803-814.

3. Bansback N, Tsuchiya A, Brazier J, Anis A. Canadian Valuation of EQ-5D Health States: Preliminary Value Set and Considerations for Future Valuation Studies. *PLoS One.* 2012 (Epub 2012 Feb 2012;7(2):e31115.

4. Robinson A, Covey J, Spencer A, Loomes G. Are some deaths worse than others? The effect of 'labelling' on people's perceptions. *Journal of Economic Psychology.* 2010;31(3):444-455.

5. Flynn TN. Using Conjoint Analysis and Choice Experiments to Estimate QALY Values Issues to Consider. *Pharmacoeconomics.* 2010;28(9):711-722.

6. 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.

7. Hakim Z, Pathak DS. Modelling the EuroQol data: A comparison of discrete choice conjoint and conditional preference modelling. *Health Econ.* Mar 1999;8(2):103-116.

8. Flynn TN, Louviere JJ, Marley AA, Coast J, Peters TJ. Rescaling quality of life values from discrete choice experiments for use as QALYs: a cautionary tale. *Population health metrics.* 2008 2008;6:6.

9. Tversky A, Sattath S, Slovic P. Contingent weighing in judgment and choice. *Psychol. Rev.* Jul 1988;95(3):371-384.

10. Robinson A, Spencer A. Exploring challenges to TTO utilities: valuing states worse than dead. *Health Econ.* 2006;15(4):393-402.

11. Keeney RL, Raiffa H. *Decisions with multiple objectives, preferences and value trade-offs*: Wiley, London; 1976.

12. Carthy T, Chilton S, Covey D, et al. On the contingent valuation of safety and the safety of contingent valuation: Part 2 - The CV/SG "chained" approach. *Journal of Risk and Uncertainty.* Dec 1998;17(3):187-213.

13. Karni E. A theory of medical decision making under uncertainty. *Journal of Risk and Uncertainty.* Aug 2009;39(1):1-16.

14. Machina MJ. Choice under uncertainty - problems solved and unsolved -responses. *J. Econ. Perspect.* Spr 1988;2(2):181-183.

15. Doctor JN, Bleichrodt H, Lin HJ. Health Utility Bias: A Systematic Review and Meta-Analytic Evaluation. *Med. Decis. Mak.* Jan-Feb 2010;30(1):58-67.

16. Abellan-Perpinan JM, Bleichrodt H, Pinto-Prades JL. The predictive validity of prospect theory versus expected utility in health utility measurement. *J. Health Econ.* Dec 2009;28(6):1039-1047.

17. de Palma A, Ben-Akiva M, Brownstone D, et al. Risk, uncertainty and discrete choice models. *Mark. Lett.* Dec 2008;19(3-4):269-285.

18. Van Houtven G, Johnson FR, Kilambi V, Hauber AB. Eliciting Benefit-Risk Preferences and Probability-Weighted Utility Using Choice-Format Conjoint Analysis. *Med. Decis. Mak.* May-Jun 2011;31(3):469-480.

19. Van Osch SMC, Stiggelbout AM. The construction of standard gamble utilities. *Health Econ.* Jan 2008;17(1):31-40.

20. van Osch SMC, van den Hout WB, Stiggelbout AM. Exploring the reference point in prospect theory: Gambles for length of life. *Med. Decis. Mak.* Jul-Aug 2006;26(4):338-346.

21. Hershey JC, Schoemaker PJH. Probability versus certainty equilvalence methods in utility measurement - are they equivalent. *Manage. Sci.* 1985;31(10):1213-1231.

22. Robinson A, Loomes G, Jones-Lee M. Visual analog scales, standard gambles, and relative risk aversion. *Med. Decis. Mak.* Jan-Feb 2001;21(1):17-27.

23. Tversky A, Kahneman D. Advances in prospect theory - cumulative representation of uncertainty. *Journal of Risk and Uncertainty.* Oct 1992;5(4):297-323.

24. van de Kuilen G, Wakker PP. The Midweight Method to Measure Attitudes Toward Risk and Ambiguity. *Manage. Sci.* Mar 2011;57(3):582-598.

25. Aitchison J, S S. The generalization of probit analysis to the case of multiple responses. *Biometrika.* 1957;44:131-140.

26. Stolk EA, Oppe M, Scalone L, Krabbe PFM. Discrete Choice Modeling for the Quantification of Health States: The Case of the EQ-5D. *Value Health.* Dec 2010;13(8):1005-1013.

27. Prosser LA, Wittenberg E. Do risk attitudes differ across domains and respondent types? *Med. Decis. Mak.* May-Jun 2007;27(3):281-287.

Appendix: The Risk Attitude questions

The mid-weight method uses a two-part elicitation process. In the first stage, two questions were used to set monetary values for £X and £Y which were equally spaced on the utility scale, in the second stage, a series of questions were used to infer the probability weighting function. In the first stage, respondents were asked to compare two risky prospects, A and B. Importantly, risky prospect A was the less attractive prospect, as it involved a chance, p, of £30 compared to £40 offered by prospect B (where p=70%). Figure 4 illustrates questions 1 and 2. In question 1 respondents faced a risky prospect B in which there was a 30% chance of winning £60 and a 70% chance of winning £40. They also faced another risky prospect A, in which there was a 30% of £X and 70% of £30. Their task was to set £X so that they were indifferent between the two options. We would expect that £X>£60 since prospect A was the less attractive prospect, involving as it did a £30 compared to £40. In question 2, risky prospect B involved a 30% chance of £X (carried over from question 1) and risky prospect A involved a 30% chance of £Y. There task this time was to set £Y so that they were indifferent between the two prospects. These two questions were designed to ensure that a movement from £60 to £X gave the same utility as a movement from £X to £Y.

In the second stage, questions were used to estimate the probability weight for different level of risk using these £X and £Y values carried over from questions 1 and 2. We estimated the probability weighting function for two points equal to 0.5 and 0.25. Participants were faced with a risky prospect B in which there was a p chance of £Y and a 1-p chance of £60. They were also faced with a prospect A that was the certainty of £X. After filling in the amounts for £X and £Y from questions 1 and 2, the respondents task here was to set p so that they were indifferent between Prospects A and B. Given that respondents had already set £X to be half way between £Y and £60 in terms of utility, we would expect that they should set p=0.5. Any divergence from p=0.5 therefore was due to their weighting of probability. Table 8 shows the pattern of risk attitude yielded by these questions.

Table 8. Probability weighting derived from risk attitude questions

	Underweight probability $w^{-1}(p)-p>0$	Equal weight $w^{-1}(p)-p=0$	Overweight probability $w^{-1}(p)-p<0$	Number*
P=0.5	36	6	12	54
P=0.25	41	2	11	54

*The pattern or responses of 6 respondents were such that risk attitude could not be determined.