Does the value of life depend on the context?

Angela Robinson[1], Judith Covey[2], Graham Loomes[1], Anne Spencer[3]

[1] **University of East Anglia**
[2] **University of Durham**
[3] **Queen Mary University, London**

The work presented here is based on a preliminary draft of a report submitted to the Health and Safety Executive and is not to be cited without the authors' permission or circulated outside the HESG.

Address for correspondence:
Angela Robinson
School of Medicine, Health Policy & Practice,
University of East Anglia,
Norwich
10603 593620

e mail: angela.robinson@uea.ac.uk

BACKGROUND

This paper reports a study conducted on behalf of the Health and Safety Executive (HSE), who, along with other government departments, make decisions about the allocation of resources across a broad range of health and safety programmes. In doing so, they face the difficult issue of whether to attach greater weight to the prevention of certain types of fatalities than others and recognise that such judgments ought to be guided by the preferences of the population affected. There is evidence to suggest that the value of preventing a fatality may vary from one hazard context to another, reflecting different degrees of dread at the prospect of death or injury in different circumstances, together with different perceptions of the degree of control, responsibility etc. And, of course, the age distribution of victims will vary across hazards and the public may wish more weight be given to certain age groups than others.

The Department for Transport (DfT), currently use a willingness to pay (WTP)-based value for preventing a fatality (VPF) on the roads of approximately £1.25 millions, but how transferable is this figure to other contexts? One means of answering this question would be to estimate directly the relevant wealth/risk trade offs across all hazard contexts in a series of contingent valuation studies. The approach taken here, however, was to seek to determine the ratio of the value of preventing one kind of hazard *relative* to another. Such a set of relative weights may then be applied to some absolute monetary 'peg' – such as the DfT's roads VPF- and the value of preventing fatalities in other contexts inferred.

We sought to gather information about these weights by asking respondents to answer in their capacity as *citizens* and express their preferences over general principles of social decision making regarding different life-saving interventions. The adoption of a citizen's perspective allows individuals to express preferences over hazard contexts to which they are not themselves exposed and to make judgments about the relative weighting that ought to be given to different age groups. An earlier study commissioned by the HSE adopted a citizens' perspective in asking respondents a series of 'matching' questions between a number of different hazard contexts (for example, roads, rail, domestic fire, fire in public places etc). Briefly, matching – in this case, person trade off (PTO) -questions ask people to state the number of outcomes of one kind they consider to be 'just as good as' a specified number of outcomes of another kind. In those questions, the number of deaths prevented took on the

role of a metric- or numeriare- in establishing the strength of preference for the prevention of one type of death over another.

There is, however, evidence (e.g. Tversky et al., 1988) that asking people to 'match' on any single dimension (for example, number of people treated) encourages respondents to attach undue weight to that specific dimension while neglecting other factors that they would otherwise wish to be taken into consideration. This may help explain why, despite expressing very strong views about the prevention of rail deaths in particular, respondents in the earlier study did not go on to attach much more weight to those deaths in the matching questions. In addition, responses were found not to be multiplicatively transitive[1] (see Chilton et al, 2002), a funding reported previously by Ubel and colleagues (1996). Elsewhere, Pinto-Prades (1997) has found marked differences between different variants of matching questions.

Thus, as an alternative, we elected to explore the use of a discrete choice (DCE) format in this later HSE study. In the present context, the attributes in a DCE design included characteristics of hazards such as the age of a typical victim, length of illness or suffering preceding death, who is to blame for the death etc. If the *specific* cause of death (for example, as a driver in a car accident) did not matter to respondents, such a 'generic' model could be estimated and used to predict the utility of life saving intervention with any particular combination of attribute levels. That is, there would exist a model into which decision makers could simply 'plug in' the relevant characteristics of any particular hazard in order to estimate the relative weight attached to preventing a fatality of that nature.

On the other hand, it may be that the specific cause of death does matter *over and above* knowledge of its generic characteristics. On the face of it, it may seem like the specific cause of death (for example, as a driver on the roads or as a passenger on the railways), may simply be added as an attribute in the choice set: but this raises particular problems for the study design. A practical problem in attempting to devise a 'context specific' DCE design is that there would be insufficient variation in levels to allow a statistically valid model to be estimated. For example, an orthogonal study design may link the attribute 'typically affecting under 17 year olds' to 'dies of lung cancer', or the attribute 'dies as a passenger in a rail

[1] That is, if a respondent equates 10 A to 20 B, and separately equates 20 B to 40 C, we might infer that she judges 10 A equivalent to 40 C: but when asked to make this trade-off directly, the ratio is often (much) less extreme.

accident' with 'the individual is mostly to blame for their own death' which are clearly implausible.

We set out to estimate a 'generic' DCE model and to explore to what extent such a generic model could predict choices made once the specific cause of death- or context- had been revealed. In doing so, our analysis raised questions about the utility function that is commonly estimated in DCE and problems relevant to the use of DCEs in health economics.

METHODS

The 'generic' DCE model

Following extensive piloting, the attributes and levels associated with those attributes were set out as follows:

- **Number** of deaths prevented: 10, 15, 25 or 50.
- **Age** group of typical victim: under 17 years, 17-40 years, 40-60 years, or over 60.
- **Duration** of illness or suffering prior to death: a few minutes, a couple of weeks, a year or two, or 3-5 years.
- **Severity** if illness or suffering prior to death: quality of life a bit worse than normal, quality of life a lot worse than normal
- Who is most responsible or to **blame** for the death: nobody in particular, the individuals themselves, other individuals, or business/government.

This results in 512 possible combinations. A reduced-form design was developed that involved a total of 64 scenarios paired to give 32 choices between two hazard scenarios. The design started with an initial set of 16 scenarios and followed a 'foldover design' which allows for main effects to be estimated and which controls for 2-way linear interactions (see Louiviere et al, 2000). Those 32 'generic' pairs were then divided between three Versions of the questionnaire, with two pairs being common across all three.

The original intention had been to present each pair in the form of a choice between two life-saving interventions, and to ask which one was 'better'. However, piloting showed that an uncomfortably large minority of respondents were liable to be confused by this question[2].

[2] The expectation had been that people would identify the scenario they regarded as worse and nominate the intervention which would prevent that scenario as the better intervention. Unfortunately, too many respondents simply considered which scenario was less unpleasant and selected that.

It was therefore decided to ask the questions in the form that people found easier to answer: that is, to describe two scenarios and ask them to say which of the two they considered to be worse, and *how much worse* they considered it to be[3]. Thus the two scenarios shown above would have appeared in the final format as follows:

Which is worse?

	A	**B**
Number of people who die	15 deaths	10 deaths
Age-group	Under 17 year olds	17-40 year olds
Quality of life in period leading up to death	A lot worse than normal for last 3-5 years of their lives	A bit worse than normal for last 1-2 years of their lives
Who is most to blame	Nobody in particular	The individuals themselves

What do YOU think?	A is <u>much</u> worse than B	A is <u>slightly</u> worse than B	B is <u>slightly</u> worse than A	B is <u>much</u> worse than A
(tick one)	☐	☐	☐	☐

In total, the generic part of the questionnaire – a booklet labeled Section 1 – consisted of 23 questions in that format. The first five of these were 'practice' questions. For each practice question the attributes of both scenarios were the same except for one item. Each of the five practice questions then varied a different attribute. Questions 6 to 23 were made up as follows. Twelve of them were part of the DCE design: two (Questions 10 and 18) were common to all three Versions and the other ten were unique to a particular Version. Splitting the questions between versions in this way gave coverage of all 32 pairs (i.e. 2 common questions and 30 version specific questions). In addition, the two common questions allowed a check of comparability across the versions.

The other six questions involved five pairs that would appear again in Section 2 with contextual information; but in Section 1 this contextual information was omitted and only the generic information was given. One of these five pairs was presented twice – first as Q6, then again as Q21 – in order to check stability of responses. These questions were not part of the DCE design and were different in each of the three Versions. These questions were included

[3] We elected to include 4, rather than 2, response categories in order to allow a more sensitive comparison of generic and contextual choices.

for two reasons. Firstly, to see how well our estimated model could predict patterns of response to other generic scenarios that were not part of the estimation procedure. Secondly, to examine whether additional contextual information lead to any significant differences.

The contextual scenarios

In Section 2, respondents were presented with another 5 pairs of scenarios, but this time with contextual information.

Question	**Version 1**	**Version 2**	**Version 3**
Q24	10 car drivers vs 10 rail passengers	15 car drivers vs 10 rail passengers	25 car drivers vs 10 rail passengers
Q25	50 smoking cancer vs 25 asbestos cancer	25 CO poisoning vs 15 accidents at work	15 smoking cancer vs 10 asbestos cancer
Q26	25 car drivers vs 15 pedestrians	10 accidents at work vs 15 car drivers	25 accidents at work vs 50 car drivers
Q27	15 pedestrians vs 25 breast cancer	25 pedestrians vs 15 breast cancer	15 CO poisoning vs 25 accidents at work
Q28	10 work-related cancer vs 15 car drivers	25 work-related cancer vs 50 car drivers	15 car drivers vs 25 pedestrians

Space does not permit a discussion of all the contextual pairings included in the study design and we limit ourselves here to looking at the car driver/rail passenger comparisons. In all three Versions Q24 paired deaths of drivers in road traffic accidents with deaths of passengers in rail accidents. What varied across the three Versions was the number of deaths: in Version 1 it was 10 deaths of both kinds, while in Version 2 it was 15 road and 10 rail, and 25 road versus 10 rail in Version 3.

Implementation

Following unsuccessful pilots of postal/telephone data collection methods, it was decided to convene a number of groups to be moderated by JC, GL and AR to administer the questionnaires. Groups were convened in the North East and East Anglia regions. A total of 313 respondents took part.

Respondents were first taken through an introductory sheet where specific examples were used to illustrate the way scenarios might correspond with real-world (albeit simplified and stylized) cases. It was emphasized that these examples were illustrative only and that the questionnaire they were about to see would be asking them to think in more general terms: i.e. in terms of general principles rather than particular cases. Section 1 was then given out and respondents taken through the five practice questions, with brief discussions after each one. They were then asked to work through the rest of the booklet themselves. When all group members had completed Section 1, there was an opportunity for a brief discussion.

Section 1 was then collected and Section 2 was handed out. Respondents' attention was drawn to the fact that these questions gave contextual information and that there was an invitation not only to tick a box but also to write a sentence or two in support of their decision. Finally, basic demographic information was elicited along with details of existing illness in the household.

RESULTS

Though not reported in detail here, the impacts of differences along each dimension, when considered in isolation in the 5 practice questions, were much as might have been expected. That is, the more people dying, the greater the pain and suffering, the longer duration of pain and suffering and the more blame rested with business or government, the worse the situation was considered to be. The test of stability of responses to the repeated question showed responses to Q6 and Q21 to be significantly different at the 5% level in Version two only.

Sensitivity to number of deaths.

If we focus on Q21, Table 1 below shows the patterns of choice in each Version of the questionnaire where A>>B indicates the response 'A is <u>much</u> worse than B', while A>B denotes 'A is <u>slightly</u> worse than B'; A<B and A<<B are the corresponding statements in the opposite direction. Comparing Versions 1-3, responses tend to migrate in the expected direction. A becomes progressively worse, as the number of deaths increases from 10 to 15 and then to 25, while the number of deaths in B stays at 10 in all three Versions. Table 1 shows the percentages making each response in the three Versions. However, it is noteworthy that although the tendency ***is*** in the right-to-left direction we should expect, it is not especially responsive to the changes in the numbers of deaths in the A scenario. Notably the B scenario

differed from the A scenario on two attributes aside from the numbers of deaths. In B the victims' quality of life was a lot worse than normal as opposed to a bit worse than normal, and business and government were to blame rather than the individuals themselves.

Table 1 Q21 in Versions 1-3
(Deaths in A Increase 10→15→25)

	A>>B	A>B	A<B	A<<B	N
Version 1	5.0	8.9	37.6	48.5	101
Version 2	3.7	13.0	53.7	29.6	108
Version 3	11.2	18.4	40.8	29.6	98

Comparing Generic with Contextual Questions

Consider Q24, which presented the same attributes as Q21, but this time labeling the deaths in scenario A as car drivers, while the deaths in scenario B were identified as rail passengers. Comparing Table 2 with Table 1, there is a more pronounced tendency for those who consider B worse than A to judge it to be much worse and a slight tendency for the A:B split to shift towards B. Even in Version 3 when there are 2.5 times as many car driver deaths, B was still considered worse or much worse than A by nearly 3 in 4 respondents (73.4%).

Table 2 Q24 in Versions 1-3
Car Drivers (A) vs Rail Passengers (B)

	A>>B	A>B	A<B	A<<B	N
Version 1	7.9	5.0	21.8	65.3	101
Version 2	4.7	9.3	37.4	48.6	107
Version 3	10.2	16.3	26.5	46.9	98

The reasons why respondents considered rail accident deaths to be worse than car accident deaths were explored by analyzing the written comments they provided. Particular attention was paid to the comments of those respondents who had switched their choice from Q21. These comments indicated that the including the context had reinforced the impact of the blame dimension, and that the public thought railways ought to be safe or that railway deaths

are avoidable. The horrific nature of a death in a rail accident was also cited as a reason for switching between the generic and contextual questions.

The generic discrete choice model

This section describes the results of the generic model estimated from responses to the 32 questions that made up the main DCE design. While we have categorical data- from A much worse than B, to B much worse than A- for the purposes of estimation using a binary logit model, we combined the categories 'much worse' and 'slightly worse' and considered only whether A was regarded as worse than B or vice versa. As there were no significant differences in responses to the two questions – Q10 and Q18 – that were common to all three questionnaires, we considered it appropriate to pool the data across the three versions of the questionnaire.

In principle, the disutility[4] of *a death of a particular type* might be expected to be some function of the age of the typical victim, the duration and severity of any period of ill-health prior to death, and the primary responsibility/blame for the death. On this basis, the disutility of *a particular scenario* would be the disutility of each death of that type *multiplied by* the numbers of deaths specified in the scenario. Thus if the disutility of a type X death is given as $U(X) = f(age_X, severity_X, duration_X, blame_X)$, a scenario involving 50 such deaths will entail a total disutility of 50U(X): that is, twice as much disutility as a scenario involving 25 deaths of that type.

However, it became apparent in the course of discussions with respondents that a proportion of them were not processing the information about numbers of deaths in that strictly multiplicative way. Rather, some respondents seemed to focus more on *differences* in the numbers, or on levels. For example, what some people perceived as sufficiently small differences – e.g. between 15 and 10 – might be given little or almost no weight, despite the fact that one was 50% greater than the other. To explore these issues in the quantitative analysis, we estimated both a difference model and a multiplicative model.

[4] It is probably more appropriate to use the term 'disutility' here as we asked respondents to identify which scenario they consider to be *worse* than the other.

The difference model

The first model is the one typically used to model discrete choice data, referred to henceforth as the difference model. Here each of the attributes, *including the number of deaths*, are taken as independently contributing to the disutility of a scenario. The model estimated is of the following general form:

$$P(B) = f\{\gamma_n(N_B-N_A) + \gamma_a(age_B-age_A) + \gamma_s(severity_B-severity_A) + \gamma_d(duration_B-duration_A) + \gamma_b(blame_B-blame_A) + e\}$$

Where P(B) is the probability that a respondent will consider scenario B to be worse than A, N_A and N_B are the number of deaths in scenarios A & B respectively, γ = the coefficient on the relevant attribute. The difference in the number of deaths, (N_B-N_A), enters the model as a continuous variable. The remaining attributes enter the model as dummy variables with the omitted dummies representing the following base case: Age = over 60's, Severity = *bit* worse than normal, Duration = for last *few minutes* of their lives, Blame = nobody in particular

Observations were not all independent of one another (since each respondent contributed 12 observations), so standard errors were adjusted to allow for clustering per respondent. The parameter estimates from the difference model and their respective levels of significance are given in Table 3.

Table 3: Model 1 – the Difference Model

Variable	Coefficient	Robust Std Error	Significance
number of deaths	.028	.002	0.000
age(<17)	1.511	.124	0.000
age(17-40)	1.080	.106	0.000
age(40-60)	.635	.090	0.000
severity(lot worse)	.624	.063	0.000
duration(weeks)	.268	.085	0.002
duration(1-2 yrs)	.446	.087	0.000
duration(3-5yrs	.596	.094	0.000
blame(individual)	-.487	.106	0.000
blame(other)	.952	.101	0.000
blame(bus/gov)	1.019	.099	0.000
Log pseudo-likelihood = -1841.0107 Number of observations = 3401			

In this model, coefficients may be interpreted as 'marginal' disutilities: that is, how much an extra unit/level of that attribute in a scenario increases the disutility of that scenario. Indeed all the coefficients are significantly different from zero and appear to have the 'correct' signs. For example, 'number of deaths' has a positive coefficient, showing (as expected) that more deaths will increase the disutility of a scenario. The dummies on age show that disutility increases as the age of the typical victim falls. Also in line with expectations, the dummy for severity 'a lot worse than normal' (compared with 'a bit worse') increases the disutility of a scenario and disutility increases with duration of suffering. The dummies relating to blame show an interesting pattern. Blame(individual) is the dummy for 'the individual themselves to blame' which, according to the model, *reduces* the disutility of a scenario relative to the base case - 'nobody in particular to blame'. On the other hand, blame(other) and blame(bus/gov) - dummies representing 'other individuals to blame' and 'business or government to blame' respectively - increase the disutility of a scenario.

The multiplicative model

While Model 1 seemed to give plausible results, it was derived from an assumption about the 'number of deaths' variable that may have corresponded with how many respondents *did* process the information, but did not correspond with how, strictly speaking, they *ought* to have made their judgments. As noted above, standard utility theory entails that they should be multiplying the disutility of a particular type of death by the number of those deaths. We therefore estimated Model 2 as follows:

$$P(B) = f\{(N^{\alpha}_{B}-N^{\alpha}_{A}) + \gamma_a(N^{\alpha}_{B*}age_B-N^{\alpha}_{A*}age_A) + \gamma_s(N^{\alpha}_{B*}severity_B-N^{\alpha}_{A*}severity_A) + \gamma_d(N^{\alpha}_{B*}duration_B-N^{\alpha}_{A*}duration_A) + \gamma_b(N^{\alpha}_{B*}blame_B-N^{\alpha}_{A*}blame_A) + e\}$$

Where N^{α}_{A} and N^{α}_{B} are the number of deaths in scenarios A & B respectively raised to power α. When α is set equal to one, all deaths are given equal weight. Values of α less than 1 indicate a declining marginal disutility of deaths, so that 50 deaths would be given less than five times the weight of 10 deaths; while values of α greater than 1 indicate an increasing marginal disutility of deaths.

The disutility of the 'base type' of death[5] was accorded a value of 1. We first estimated the multiplicative model with $\alpha = 1$. We should expect the coefficients to have the same signs as in Model 1. However, a number of the parameter estimates had unexpected signs and did not appear to fit the data at all well (the value of the log-likelihood fell from a value of -1841 in Model 1 to a value of -6681 in Model 2 with $\alpha = 1$). It appeared that setting $\alpha = 1$ imposed a structure on the weight placed on the number of deaths which diverged from what respondents actually did to an extent that it distorted many of the parameter estimates.

We then explored other values of α. A grid-search showed that the log-likelihood function was minimized (i.e. the multiplicative model fitted best) when $\alpha = 0.2$. The results of estimating Model 2 on this basis are shown in Table 4.

Table 4: Model 2 – the Multiplicative Model with $\alpha = 0.2$

Variable	**Coefficient**	**Robust Std Error**	**Significance**
Deaths	(offset)		
Dage(<17)	.817	.065	0.000
Dage(17-40)	.605	.055	0.000
Dage(40-60)	.355	.049	0.000
Dsev(lot)	.331	.035	0.000
Ddur(weeks)	.152	.045	0.001
Ddur(1-2 yrs)	.276	.047	0.000
Ddur(3-5yrs	.354	.053	0.000
Dblame(indiv)	-.259	.057	0.000
Dblame(other)	.526	.054	0.000
Dblame(bus/gov)	.537	.051	0.000
Log pseudo-likelihood = -1839.4199 Number of observations = 3401			

The deaths variable is 'offset' to make the disutility of the base type death equal to 1. The variables have the same interpretation as in model one, but are prefixed by a 'D' to indicate that they are now multiplicative in the number of deaths. Notice that with $\alpha = 0.2$ the

[5] A death of someone over 60 for which nobody in particular is to blame and where the quality of life is a bit worse than normal for a few minutes prior to death.

coefficients are all signed in accordance with prior expectations. Furthermore, the log-likelihood is very similar to that for Model 1.

Although setting $\alpha = 0.2$ gave the best fit to the data used for estimation (i.e. the responses to the 32 questions in the main design), it did not always appear to be such a good predictor of responses to the other questions which involved generic descriptions of the contextual scenarios. In particular, it tended to somewhat underestimate actual sensitivity in those questions to the numbers of deaths. For this reason, we increased the value of α to 0.3 and re-estimated the model[6], with the results shown in Table 5.

Table 5: Model 3 – the Multiplicative Model with $\alpha = 0.3$

Variable	Coefficient	Robust Std Error	Significance
Deaths	Offset		
Dage(<17)	.526	.051	0.000
Dage(17-40)	.293	.041	0.000
Dage(40-60)	.121	.032	0.000
Dsev(lot)	.196	.027	0.000
Ddur(weeks)	.096	.034	0.004
Ddur(1-2 yrs)	.175	.040	0.000
Ddur(3-5yrs	.316	.051	0.000
Dblame(indiv)	-.231	.047	0.000
Dblame(other)	.295	.043	0.000
Dblame(bus/gov)	.345	.036	0.000
Log pseudo-likelihood = -1886.1191 Number of observations = 3401			

Testing the predictive power of the models.

We turn now to how well each model predicts responses to those questions that were not used in the estimation process i.e. the contextual scenarios and their generic equivalents. Again focusing on responses to Q21 (the generic equivalent of the car driver/rail passenger comparison), the first three columns of Table 6 show the predictions about the percentages of respondent who will consider B to be worse than A made by Models 1, 2 and 3 respectively .

[6] Increasing the value of α to 0.4 resulted in some coefficients having the wrong signs and/or key variables no longer being significant.

The last two columns show the actual percentage of respondents who answered in that manner. It is clear that all three sets of predictions move in the right direction. Models 1 and 2, however, predict *less sensitivity* to the numbers of deaths than found in the actual data. For example, in moving from Version 1 (10 deaths in A) to Version 3 (25 deaths in A), Model 2 predicts that the percentage of respondents rating B worse than A falls from 85.6 to 79.5, while the actual percentage of respondents who answered in that way fell from 86.1 in Version 1 to 70.4 in Version 3. Model 3 does the best job of predicting these percentages, although it still falls several percentage points short of the actual outcome in Version 1.

Table 6: Comparison of Model Predictions with Actual Responses for Q21 in Versions 1-3 (Deaths in A Increase 10→15→25)

	Model 1	Model 2	Model 3	Actual % Rating B Worse (A<B *or* A<<B)
Version 1	89.4	85.6	82.4	86.1
Version 2	88.0	83.3	78.0	83.3
Version 3	84.7	79.5	70.5	70.4

Now consider Q24, which presented the same attributes as Q21, but labeled the deaths in scenario A as car drivers, while the deaths in scenario B were identified as rail passengers. Table 7 again shows the predicted percentages and the actual responses to that question.

Table 7: Comparison of Model Predictions with Actual Responses for Q24 in Versions 1-3 (Deaths in A Increase 10→15→25)

	Model 1	Model 2	Model 3	Actual % Rating B Worse (A<B *or* A<<B)
Version 1	89.4	85.6	82.4	87.1
Version 2	88.0	83.3	78.0	86.0
Version 3	84.7	79.5	70.5	73.4

In this case, all models underestimated the sensitivity of responses to changes in the numbers of deaths of car passengers. Model 2 fitted marginally better than Model 3, but both were quite a distance away from the actual responses in Version 1. When those results that are not

reported here are included, Model 1 performed rather worse than either of the other two, being much more likely to underestimate the sensitivity of responses to changes in the numbers of deaths and also being generally poorer at predicting the actual percentages.

COMPUTING RELATIVE WEIGHTS

The major drawback of Model 1, which is based on absolute differences in number of deaths, is that it cannot produce stable ratios of the type sought here. For example, suppose we set the number of type X deaths at 10 and used the model to compute the number of type Y deaths that would be required to make both scenarios equal in terms of disutility. Suppose that were achieved when Y = 30. It would then appear that each type X death was equivalent to 3 type Y deaths. But suppose instead we set the number of type X deaths at 100 and repeat the process, calculating that the equivalent number of Y deaths is 120. This calculation would suggest that each type X death was equivalent to 1.2 type Y deaths, which would have very different implications for policy. Since there is unlikely to be any obvious correct base from which to start, this is an undesirable property of using a model which involves using differences between the numbers of deaths in the estimation procedure.

Given that the estimation of this model involves something that is theoretically suspect – i.e. taking the number of deaths as an additively separable variable- we cannot recommend it as the basis for policy judgments, even though the differences between the numbers of deaths may have been what a proportion of respondents were focusing on. The multiplicative specification is rather easier to defend. It still assumes that the characteristics which contribute to the disutility of any particular *type* of death do so in an additively separable way, which may not necessarily always be the case. But by multiplying types of deaths by the numbers of those deaths, this specification treats the 'number of deaths' variable in the appropriate manner and enables us to estimate ratios which are independent of the particular numbers chosen and which are transitive. Thus relativities may be computed from some base case and the weights derived will be independent of the particular numbers used in that base case.

For the multiplicative models, it is possible to vary the ratio N_A:N_B until any two scenarios are predicted to be equal in terms of disutility, in the sense that the probability of B being rated worse than A becomes 0.5. In principle, this ratio provides the basis for the relativity weights.

However, there are substantial – and, arguably, rather uncomfortable – implications of setting α equal to 0.2 or 0.3. In particular, rather than 50 deaths being five times as bad as 10 deaths, setting $\alpha = 0.3$ implies that respondents regard 50 deaths as being just 1.62 times as bad as 10 deaths[7]. This marked insensitivity to the numbers of deaths clearly impacts on the weights derived.

Using the example of the car driver/rail passenger comparison, Model 3 suggests that the number of deaths of car drivers that would be regarded as equally as bad as 10 rail passenger deaths is approximately 62, suggesting a ratio of 6.2:1[8]. Translating this into the value of preventing a fatality, if the VPF associated with a car driver were £1.25m, the corresponding VPF for a rail passenger would be £7.75m[9].

An alternative would be to use the coefficients on the various dummies estimated from Model 3 to calculate the disutility of any particular type of death as determined by the set of characteristics of that death, and compute relativities on that basis. For example, the disutility score for a car driver death as described in the study would be 1.062 while the score for a rail passenger death would be 1.834. This would give a ratio of 1.727:1, which would translate into a VPF for a rail passenger of £2.16m. Clearly, the latter procedure greatly compresses the differential.

The question is: given the substantial difference between the various ways of calculating ratios, how should we decide where to locate the ratios upon which policy might be based? Is it appropriate for policy decisions to reflect diminishing marginal utility over the numbers lives saved even if that reflects the public preference? Or is such insensitivity an artifact of the elicitation method and, as such, something that ought to be 'factored out' of policy

[7] In this case the ratio is 5, and $5^{\alpha} = 1.62$ when $\alpha = 0.3$.

[8] It is worth stressing that the same ratio is derived for *any* starting value of rail passenger deaths. In contrast, the difference model yields ratios of 8.6: 1 or 9.6:1 depending on whether the (arbitrary) starting point of 10 or 20 rail passenger deaths is used.

[9] This figure is sensitive to the way in which each type of death is characterized. In the study, the car driver deaths were described in terms of 'the individual themselves' being to blame. But this is clearly not appropriate in the case of *all* car drivers. Thus, the ratio of 6.2:1 is likely to be an overestimate of the weight attached to rail passengers vis-à-vis the *average* car driver fatality. For example, if the 'blame' attribute were changed to 'nobody in particular' in the case of car drivers, the number of car driver deaths required to make that scenario equally as bad as the deaths of 10 rail passengers falls to 32, giving a 3.2:1 ratio. The ratio falls further to 1.6:1 if 'other individuals' are seen as the cause of the car driver deaths.

decisions? There does not seem to be any obviously 'right' answer to this question and we welcome comments from HESG members on this.

DISCUSSION

The use of DCEs in health economics has increased significantly in recent years (Ryan and Gerard, 2003) and applications now include the estimation of utility values and QALYs (Viney & Savage, 2003, Bryan et al, 2002, Viney et al ,2005). This has resulted in the design of DCEs that include attributes such as; health status, probability of survival, duration of survival and numbers of people treated. Although the resulting choice problems are fundamentally different than those to which DCEs have traditionally been applied, relatively little attention has been paid to the appropriate functional form of the model. We believe this has resulted in the modeling of attributes that are essentially *multiplicative* as if they were additive.

It is important to note that the 'Number of deaths' variable was included here as a scalar- or numeriare – in order to assess the relative strength of preference over types of deaths and, as such, fulfills the same role as in a matching or PTO question. Similarly, variables such as 'probability of survival' or 'duration of survival' play the roles of scalars in standard gamble (SG) and time trade off (TTO) studies respectively.

In one study by Bryan et al which set out to test the QALY assumptions using a DCE, respondents were presented with various health programmes and asked to make a series of choices (Bryan et al, 2002). Four attributes were included in the design: the number of people benefiting from treatment, the chance of success of treatment, duration of survival and quality of survival and the model based on absolute differences between levels on attributes. Yet, it us unclear why probability of success, numbers of people treated, duration of survival and quality of survival ought to be modeled as independent variables or why respondents ought to consider absolute differences across levels on each. The number of people who benefit from treatment is clearly the product of the number of people treated and the chances of success of that treatment. The amount of benefit each person receives is then presumably some combination of quality and duration of survival. As in our study, we believe the underlying preference structure ought to be multiplicative in nature

In a paper published recently in Health Economics, Viney and colleagues report a DCE designed to be an analogue of the SG and TTO tasks (Viney et al, 2005). Respondents were asked to choose between a 'treatment' and a 'no treatment' option and attributes included: out of pocket cost, chances of survival with treatment, chances of survival without treatment , life expectancy with treatment, life expectancy without treatment, health status. Where survival with treatment is in EQ-5D state 11111 whilst survival with out treatment is in one of 4 other health -H1 to H4.

Probability of survival, duration of survival and health status appear independently in the 'no treatment' arm even though two of the health states H1- representing EQ-5D state 33333 -and H2- representing EQ-5D state 22323 are likely to be considered as worse than dead by a number of respondents[10]. For states considered to be better than dead, increased probability of survival and duration of survival would both be expected to *increase* the utility of an option but the reverse is true for states considered to be worse than dead. Respondents will prefer states they consider to be worse than dead to be accompanied by *lower* probabilities of survival and *fewer* life years[11]. As preferences over the direction of one attribute depends on the level of another, it seems to make little sense to separate them out in this manner.

Again, we believe that choices ought to be modeled as if they were multiplicative. It is important to note, however, that using a multiplicative specification does not, in itself, impose QALY restrictions on the model as suggested elsewhere (Viney & Savage, 2003). Consider a general form of a multiplicative equation:

$$Uij = Pj^{\alpha} * N^{\beta} * [Tj^{\delta} * HS] + Cj$$

Where Pj is chances of success, N is the number of people treated, T is time in health state, HS is health status, C is cost and α, β, δ,are weighting functions allowing for non-linearities with respect to chances of success, numbers treated and time in health state respectively[12]. Setting $\alpha = -1$, would impose the EU specification of linearity in probabilities. Setting $\beta = 1$ would impose the QALY assumption that treating 100 people is twice as good as treating 50.

[10] The MVH TTO tariff means for H1 and H2 were -0.594 and 0.024 respectively.
[11] It is important to note that this general point does not rely on the imposition of the QALY assumption that utility is increasing (decreasing) *monotonically* with duration for states considered to be better (worse) than dead..
[12] It may also be the case that there are non-linearities with respect to health status and cost too.

As in this study, a value of $\beta < 1$ would imply diminishing marginal utility of numbers treated[13]. Setting $\delta = 1$ would impose the QALY assumption that 10 years in health state HS generates twice as many QALYs as 5 years in that state (presuming that HS is better than dead). A value of $\delta < 1$ would imply diminishing marginal utility with respect to duration such that 10 years in that state is less than twice as good as five years: the lower the value of δ, the more marked is the non-linearity. A value of $\delta > 1$ that 10 years in health state HS is more than twice as good as 5 years: which may be the situation whenever there has been adaptation over time. And, of course, interactions between health status and time are necessarily captured within this specification too[14].

We cannot see how estimating a model based on differences across attributes that combine in an additively separable way can be used to test such a specification, yet there seems to be a general view that the results of differenced DCEs may be used to test QALY-type restrictions. There is also some suggestion that, as the RUT framework is a very general one, DCE models may be estimated without the need to impose *any* specific functional form on the data. For example, in an earlier write up of the study published recently in Health Economics, Viney and colleagues set up the general specification of the utility function for the ith consumer as:
Uij = U(Pj, HSj Tj C j)

They go on to say: 'This general utility function allows for alternative- specific utility functions and does not impose any restrictions on the form of the utility function. Within this general form, the restrictions of the QALY model can be tested' (page 16, Viney & Savage, 2003) Whilst we agree that the general specification given above does not impose any restrictions on the utility function, neither does it provide us with a means of estimating a specific model. Some functional form must have been imposed on the deterministic component of the model reproduced in the recent Health Economics paper, but we are not told what that was (although the stochastic component is described in detail). This rather gives the impression that the deterministic component is irrelevant, but that is clearly not the case.

It seems reasonable for the authors to point to their results as evidence that the 'correct' models were estimated. In the Viney et al study the coefficients on the variables are reported

[13] And, as in our study, it is an open question whether decision makers would wish to take account of such diminishing marginal utility over numbers treated.

to be highly significant and in the 'right' direction as so often appears to be the case in DCE studies. Similarly, the Bryan et al paper reports that 'the coefficients relating to all four attributes were highly significant indicating that, in general, the choices respondents made were sensitive to variation in the levels for these factors'. Likewise, the coefficients in our difference model were highly significant and in the right direction: but does this necessarily mean that this was the 'correct' model to set up in the first place?

We have argued that certain attributes must be multiplicative in nature. Take, for example, probability- there is simply nothing else one may do with a probability than to multiply an outcome through by it (or some transformation of it)- yet the result of DCEs appear to show this as an additively separable variable. One possible explanation is that the format of DCE questionnaires leads respondents to treat variables that are multiplicative as if they were separable. That is, it may be the procedure of presenting respondents with levels on attributes side by side that encourages them to treat them independently even when it makes no sense to do so. After all, we are taught to read from left to right across the page starting on the first line, so this seems a reasonable thing to do when we are presented with an unusual task. Treating all attributes as if they were separable may be a simplifying strategy then used to process information that actually involves complex relationships between variables. Whilst this strategy makes it more likely that DCE models will produce 'good' results (in the sense of highly significant coefficients and signs in the expected direction), there is perhaps a danger that the estimated parameters are to an uncomfortably large extent artifacts of the procedure rather than reflections of the values and preferences we seek to elicit.

[14] In this way, the multiplicative model here is quite different from QALY calculations that take a utility value estimated outside the model then multiply through by time spent in that state.

ACKNOWLEDGEMENTS

We are indebted to Brett Day for his help with the econometric analysis presented and to Peter Moffatt for providing additional econometric advice. The idea of estimating a multiplicative model grew out of discussions with Bob Sugden and Ian Bateman and we acknowledge their considerable input into the ideas expressed here. Jordan Louiviere, Deborah Street and Emma McKintosh offered useful advice on the design of DCE experiments and we are grateful to them.

REFERENCES

Bryan, S. Roberts, T. Heginbotham, C and McCallum, A . QALY-maximisation and public preferences: results from a general population survey. *Health Economics* (2002): 11 : 679-693.

Chilton S, Covey J, Hopkins L, Jones-Lee M, Loomes G, Pidgeon N and Spencer A (2002) Public perceptions of risk and preference-based values of safety. *Journal of Risk and Uncertainty* 25(3): 211-232.

Louviere, J, Hensher, D, Swiat, J, (2000) Stated Choice Methods : Analysis and application, Cambridge University Press,

Pinto-Prades J-L (1997) Is the person trade-off a valid method for allocating health care resources? *Health Economics*, 6(1): 71-81.

Ryan M and Gerard K (2003) Using discrete choice experiments to value health care: current practice and future prospects. *Applied Health Economics and Policy Analysis* 2: 55-64.

Sloane N, A library of over 200 orthogonal arrays: www.research.att.com/~njas/).

Tversky A, Sattath and Slovic P (1988) Contingent Weighting in Judgment and Choice. *Psychological Review* 95: 371-84.

Viney, S and Savage, E. Modelling preferences for health care . Paper presented to the Labour Econometrics Workshop in Melbourne 2003.

Viney, S. Savage, E and Louviere J. Empirical investigation of experimental design properties of discrete choice experiments in health care, *Health Economics* (2005), in press

Ubel PA, Richardson, J and Baron J (2002) Exploring the role of order effects in person trade-off elicitations. *Health Policy*, 61: 189-199.