TESTS OF UTILITY INDEPENDENCE IN THE QALY MODEL WHEN HEALTH VARIES OVER TIME: TWO EXPERIMENTAL STUDIES.

ANNE SPENCER [1] & ANGELA ROBINSON[2]

([1] Department of Economics, Queen Mary College, London & [2] Centre for the Economic and Behavioural Analysis of Risk and Decision, University of East Anglia)

Prepared for CES/HESG meeting, Jan 2004.

Work in Progress- please do not quote without authors' permission.

INTRODUCTION

Background

The QALY approach combines a measure of a respondent's preferences for time and for the constituent health states. Consider three health states X, Y and Z which make up a health profile which we denote as XYZ. The QALY approach assumes that it is valid to estimate the utility of the health profile XYZ by simply adding the utilities of its constituent health states, appropriately weighted by a measure of a respondent's preferences for time (represented by w_i). The QALY approach applies, therefore, an additive model and the utility of profile XYZ is estimated by equation (1).

$$U(XYZ) = w_1 \; U(X) + w_2 \; U(Y) + w_3 \; U(Z) \qquad (1)$$

where w_i is the time discount factor at time i, for i=1,2,3 and U(.) is the utility function.

When health varies over time it is important to test whether preferences are unaffected by the order of health states. A challenge to the QALY approach arises from concerns that respondents may have preferences over the ordering of events, known as sequencing effects (Gafni, 1995; Ross & Simonson, 1991). A respondent may desire to overcome ill-health and look forward to good health (dread and savouring, Loewenstein & Prelec, 1993). A respondent may also pay more attention to the final health state in a treatment (Kahneman, at al 1993; Varey & Kahneman 1992) or be aware that they will adapt to health in a positive or negative manner over time (Ross & Simonson, 1991).

Keeney and Raiffa (1976) distinguish three cases when preferences are unaffected when health states vary over time:

1) Preferential independence (riskless choices)

Preferential independence holds if preferences between profiles that contain the same health state in period i do not depend upon the severity of the health state in period i (Keeney & Raiffa, 1976, p.101).

2) Utility independence (risky choices)

Utility independence holds if preferences between Paired Gambles (PGs) that contain the same health state in period i do not depend upon the severity of the health state in period i.

3) Additive independence (risky choices)

Additive independence holds if the preferences between risky treatments depend only upon the marginal rather than the joint probability distributions of the health states (Bleichrodt & Quiggin, 1997, p.154; Keeney & Raiffa, 1976, p.230). When health varies over time, Bleichrodt (1995) and Bleichrodt & Quiggin (1997) have shown that for QALYs to be a valid measure under Expected Utility Theory, it is necessary to assume that additive independence holds.

Results of previous studies

In testing preferential independence, Treadwell (1998) asked respondents to choose between two profiles that occurred with certainty and included health state Z in period i. He then tested whether changing the severity of health state Z altered a respondent's choice between these two profiles. Given that the comparison of health states was made within the same period, this test offers a simple technique to control for a respondent's preferences for time. Preferential independence was found to hold in 36 out of the 42 tests.

Spencer (2003) used two tests. The first test investigated additive independence using a paired gamble question. The test is based on two profiles that contain health state Z in period i and checks whether changing the severity of health state Z in period i alters a respondent's preferences between these two profiles. For example, in version 1 of this test, she compares the profiles ZNN and ZYZ with the profiles NNN and NYZ. The null hypothesis of the test predicts that additive independence holds and the differences in the SG utilities between profiles ZNN and ZYZ are the same as the differences in the SG utilities between profiles NNN and NYZ. The second test investigated the implications of the additive model under uncertainty but did not test utility independence since the test did not use a PG question. In the first test few respondents were consistent with additive independence. In the second test, only one of the two versions of the tests detected statistically significant differences.

In addition, a number of studies have looked at whether health quality and survival duration are utility independent when health is *constant* over time. For example, Miyamoto and Eraker (1988) use a SG question and offer patients a choice between remaining in a heath state X, for say t years, or undertaking a risky treatment. The risky treatment offers 50% chance that it would result in a longer period in health state X and a 50% chance that it would result in shorter period in the same health state X. The patient is asked to set time t, so that they do not mind which treatment they receive. If utility independence holds, time t should be the same for questions involving a health state X which is equivalent to a patient's own health or normal health. Miyamoto and Eraker conclude that survival duration is utility independent of quality of life.

Bleichrodt and Johannesson (1997) use a more conventional SG question, where respondents are offered a choice between remaining in health state X, or undertaking a risky treatment with a probability of normal health and death. They test whether quality is utility independent of survival duration by varying survival duration across different SG questions and seeing whether this has an impact upon the utility score that they derive for health state X. There should be no effect if quality is utility independent of survival duration. They found evidence that health quality is not utility independent of survival duration. As far as the authors are aware, utility independence has not been tested when health varies over time and this formed the basis of the first of our two studies.

STUDY ONE

The aims and objectives of the first study were;

- To carry out a quantitative test of utility independence in risky choices for mild and severe health states.
- To test the impact of changing health at the beginning or end of life.
- To explore the factors affecting decisions over risky choices from a list of predefined factors.

Methods

Keeney and Raiffa suggested a paired gamble type approach to testing utility independence but, after piloting, we opted to use the simpler 'standard gamble' format. Basically, independence means that if a respondent is indifferent between the certainty of profile X,Z,Z and a p% chance of profile X,X,X and 1-p% chance of X,Y,Y they should also be indifferent between the certainty of N,Z,Z and a p% chance of N,X,X and 1-p chance of N,Y,Y. In other words, preferences over the risky choice ought to be independent of the severity to an element that is common throughout.

A set of 'life profiles' were developed each covering the last 25 years of life, made up of 5 periods of five years. Four states were used in these profiles and were colour-coded such that normal health (N) was represented by pink, mild disability (Y) by yellow, severe disability (B) by blue, and death (D) by black. In the notation below NNNBB denotes 15 years in normal, followed by 10 in blue, whereas YNNNN denotes 5 years in the yellow state followed by 20 years in normal health.

Respondents were first asked a 'practice' SG question in order to familiarise them with the response format. This question asked them to compare a gamble with NNNNN (25 years in normal health) as the best outcome and DDDDD (death 25 years early) as the worst outcome, to the certainty of NNDDD (10 years in normal health followed by death). Respondents were then presented with a range of chances of success and failure associated with the gamble and asked to consider whether they preferred the certainty, preferred the gamble, or found it too hard to choose between those two options.

After completing the practice question, respondents were presented with 5 tests of utility independence, each test comprising of two SG questions, A and B, making 10 SG questions in all. In each case, the two 'halves' of the independence test were answered consecutively. This was done in order to minimise the possibility that any differences detected between the two treatments were due to 'random noise' or 'learning effects'. Table 1 details the 5 tests of utility independence explored in this study.

Table 1: The 5 tests of utility independence in study one.

		Better outcome P%	Worse outcome (1-P) %	Certainty of:
Test one	1A	NNNN**N**	BBBB**N**	NNBB**N**
	1B	NNNN**B**	BBBB**B**	NNBB**B**
Test two	2A	**N**NNNN	**N**BBBB	**N**NNBB
	2B	**B**NNNN	**B**BBBB	**B**NNBB
Test three	3A	**N**NNNN	**N**YYYY	**N**NNYY
	3B	**Y**NNNN	**Y**YYYY	**Y**NNYY
Test four	4A	**NN**NNN	**NN**DDD	**NN**YYY
	4B	**BB**NNN	**BB**DDD	**BB**YYY
Test five	5A	NNNN**N**	BBBB**N**	NNNB**N**
	5B	NNNN**B**	BBBB**B**	NNNB**B**

Tests one and five explore the impact of changing the health state in the last period from normal health to the severe state B. Note that the only difference between tests one and five is that the former has an extra five years of severe health associated with the certain outcome. This was done in order to test whether the duration of the period of severe health offered with certainty affected choices over the gamble. As the best and worst outcomes associated with the gamble in tests one and five are identical, we would expect the indifference probability in 1A to be lower than in 5A, and that in 1B to be lower than in 5B.

Tests two and three explore the impact of changing the health states in the first period from normal health to severe disability (test 2) or mild disability (test 3). Test four examines whether preferences over mild disability for sure vis a vis normal health and death at the end of the profile are independent of the preceding 10 years. It seemed plausible that introducing the prospect of premature death in the worst outcome would make violations of independence more likely.

After completing their SG booklet, respondents were shown a number of general statements relating to their preferences over sequencing of health states. These included questions on whether they preferred to delay periods of ill health or whether they felt states would become more tolerable over time. As it seemed plausible that preferences over the timing of health states may be dependent on the *severity* of that state, respondents were asked to consider the blue (severe) and yellow (moderate) health states in turn.

Results – study one

The sample comprised of 64 respondents, 37 males, 27 females with a mean age of 21. The results of paired t tests comparing the two 'halves' A and B of each test are given in table 2[1].

Table 2: Matched sample paired t- tests (A vs B)

	Obs	t value	P
Test one	62	-1.160	0.251
Test two	63	0.677	0.501
Test three	63	-1.559	0.124
Test four	63	-0.196	0.985
Test five	61	3.659	0.005

Clearly, there is no significant difference between responses to parts A and B in the case of four of the five independence tests carried out. Hence, we have to conclude that utility independence generally holds in the way we set out to examine it here. It is only in test five that we do find a significant difference, in particular, a significantly greater number of respondents set the indifference value of p higher in question 5A than in 5B. This finding is slightly puzzling as test five was identical to test one other than the duration of severe health under the certain outcome. We return to this in the discussion.

We turn now to the results of the general questions explore the preferences over sequencing and adaptation etc.

[1] Though not shown here, we also carried out matched pairs sign tests, with broadly similar results. The data were also re-examined after 14 'inconsistent' respondents had been removed (those respondents who set the indifference value of p to be *higher in* test 1 than in test 5) with no change in the results.

Table 3: Results of 'agree/disagree' questions

Question	Agree	Disagree	Unsure
1. Prefer Blue early in sequence	16	38	9
2. Prefer Blue at end of life	12	38	12
3. Blue becomes *more* tolerable with time	10	31	22
4. Blue becomes *less* tolerable with time	30	10	23
5. Prefer Yellow early in sequence	23	28	12
6. Prefer Yellow at end of life	38	15	10
7. Yellow becomes *more* tolerable	23	14	26
8. Yellow becomes *less* tolerable	18	18	27

If there were no sequencing effects and discounting of health states was strictly positive, we would expect respondents to agree with questions 2 and 6 and to disagree with questions 1 and 5. Whilst the majority of responses do concur with these 'conventional' answers, there is a significant minority giving different responses. For example, of the 51 expressing a clear preference in question 5, 23 agreed that they would prefer to have the yellow health state early in the sequence, presumably due to a desire to 'get it over with'. Even in the case of the severe health state, 16 respondents agreed that they would rather it occur early in the sequence.

If notions of 'adaptation' or 'duration' did not matter to respondents, we would expect them to disagree with questions 3, 4, 7 & 8, whilst significant numbers are in agreement. For example, 30 respondents thought that the blue state would become *less* tolerable through time whilst 23 felt that the yellow health state would become *more* tolerable through time. Thus, there is some reason to believe that sequencing does matter, at least to a subset of respondents, and that the affect it has depends on the severity of the health state.

Explanation of results

In study one, we set out to see whether independence holds under conditions of uncertainty, what K&R termed 'utility independence'. We found utility independence to hold in the majority of cases examined here. There exists, however, a significant body of experimental work that suggests that the sequencing of health over time *does* matter, a notion backed up by the responses to the agree/disagree questions outlined above. For example, Loewenstein and Prelec (1993) found that violations of additive separability were caused by a desire to spread good outcomes evenly over time and that respondents prefer utility levels to improve over time. In addition, there is evidence that respondents have a 'maximum endurable time' (MET) that respondents will tolerate a severe health state, above which they find death preferable (Stalmeier et al,1996, Dolan & Stalmeier,2003).

We identify a number of possible reasons why the results of our study appear to contradict evidence from elsewhere. First, the task respondents undertook was fairly complex and they may have adopted 'simplifying strategies' in order to get through. For example, respondents may have 'edited out' information that was common across choices in order to simplify the task, making violations of independence less likely. Whilst we were keen to present the health states as occurring towards the *end of life* as we considered that to be more plausible, this meant that respondents were considering scenarios that were still a long way off . It is plausible that this may have diminished the impact of duration and sequencing.

STUDY TWO

In study two, we set out to test for the effect of duration and in a much simpler format than that used in study one. To this end, we sought to;

- Elicit preferences over life profiles under a situation of certainty using a simple ranking procedure.
- Reduce the complexity of the scenarios by omitting the mild health state and shortening the sequence to 4 x 5 year periods.
- Make decisions over the profiles more 'immediate' by bringing them up to the *next* 20 years of life.

In study two, we were particularly interested in exploring the phenomenon of MET preferences using a ranking procedure. Of particular interest is the finding that for a health state in which a subject does not want to live longer than a specified amount of time, subjects' TTO responses do not reflect this, and longer durations of ill-health are equated proportionally with longer durations of healthy life years (Stalmeier et al,1996, Dolan & Stalmeier, 2003). This results in a 'preference reversal' of the type shown below;

10 yrs migraine = 8 yrs healthy

20 yrs migraine = 16 yrs healthy

in the TTO, but,

10yrs migraine > 20 yrs migraine.

in a direct choice.

This led the authors to argue that such preference reversals are due to a 'proportional heuristic' being used in the TTO. Whilst the 'proportional heuristic' idea is one possible explanation, we believe there are others that are equally plausible. First, the series of pair-wise choices involving two TTO questions (which are essentially 'matching' procedures) and a direct choice over different durations of ill health, may result in respondents focusing on different features of the decision in the different tasks. Although respondents are told that the years with migraine will be followed by death, it is also plausible that they lost sight of the fact that the shorter duration of ill health is associated with 10 years loss of life expectancy. Thus, we were keen to examine the issue when all scenarios are assessed simultaneously and when the number of years life lost is made explicit in each case.

Methods

Two sets of 10 'life profiles', each covering the next 20 years of life, were developed and printed on small strips. The profiles represented some combination of normal health, severe impairment and death in 5 year time periods and are given in table 7 along with their associated code letters. The severe impairment - blue state - used in study 2 was EQ-5D state 23323. Whilst seven of the profiles were common to both

sets, those shown in bold highlight the differences between the two sets[2].

Table 4: The two sets of profiles used in study two.

	Set One	Set Two
1.	NNNN (A0)*	NNNN (A0)
2.	NNND (A5)	NNND (A5)
3.	NNDD (A10)	NNDD (A10)
4.	NDDD (A15)	NDDD (A15)
5.	DDDD (A20)	DDDD (A20)
6.	**NNNB (B5)**	**BDDD (BD5)**
7.	**NNBB (B10)**	**BBDD (BD10)**
8.	**NBBB (B15)**	**BBBD (BD15)**
9.	BBBB (B20)	BBBB (B20)
10.	BBNN (WD10)	BBNN- WD10

* It is important to note that in our notation for 'AX' , the X refers to the number of years dead , not the number of years spent in normal health.

Respondents were randomised to receive either set one or set two first. The 10 profiles were then shuffled and respondents asked to rank the profiles from best to worst, with ties allowed. Following piloting, we used a 'choose the best' procedure whereby respondents were first asked to choose their most preferred option and place that furthest away from them. They were then asked to choose their most preferred from the remaining set and so on, until the ranking was complete[3]. After checking and recording this ranking, they were then asked to consider certain of the profiles in more detail. The procedure which followed depended on which set they were ranking.

Set one

In set one they were first asked to consider the ranking of NNBB (B10) in relation to the 'A' cards - i.e. those representing some combination of normal health and death.

[2] The purpose of developing two sets was not primarily to look for differences between the two, but to explore different ways of using a ranking procedure to elicit values for health profiles.
[3] It was not clear, however, that respondents were actually following this procedure when ranking the profiles and a number of different 'strategies' appear to have been used.

For example, suppose that a respondent ranked B10 below NNND (A5), but above NNDD (A10), these three cards would then be pulled to one side and the respondent asked to consider this subset in more detail. They would then be given four more 'A' (normal health and death) cards - in this case A6 to A9 - that increase the years dead by increments of one year. Respondents were then asked to 'slot in' these additional cards into their sub-ranking running from A5 through B10 to A10. This was done in order to obtain a more accurate valuation of B10. This procedure was then repeated for B5, using the appropriate set of 'A cards' according to B5's position in the initial ranking. This allowed us to test for proportionality; if proportionality holds and the respondent is indifferent between B10 and A8, for example, then they will also be indifferent between B5 and A4.

If there is no MET present, and B5 is ranked higher than A5 (i.e. B5 > A5), then we would also expect that; B10 >A10, B15 >A15 and B20 > A20, indicating that all durations of blue are better than dead. By the same token, if B5 < A5, then we would expect that; B10 < A10, B15< A15 and B20 < A20, indicating that all durations of blue are worse than dead. On the other hand, the following pattern in the ranking would indicate that MET had occurred somewhere between 5 and 10 years in the severe blue health state; B5>A5, B10<A10, B15<A15 and B20<A20.

It is worth noting here, however, that we are unlikely to detect preference reversals of the sort discussed by Dolan & Stalmeier (2003) in set one. Such a reversal would entail that, for example, B10 be set equal to A8 years, B5 to be set equal to A4[4], but then B5 > B10. As shorter periods of blue are associated with *longer* periods in normal health in set one, it would be irrational for a respondent to rank B5 above B10, irrespective of how bad they consider the blue state to be.

Turning to the derivation of utility values, figures 1 & 2 show how utility indices are calculated in set one and show that worse than dead scores are assessed in exactly the same manner as better than dead scores. We consider this to be a major advantage of this approach over 'conventional' TTO methods that rely on fundamentally different

[4] This is assuming proportionality, but that is not necessary for the preference reversal to occur; B5 set equal to any 'A' card lower than A8 would demonstrate the phenomenon.

procedures to assess states better than and worse than dead. In the case of B10, the worse than dead scores automatically lie between 0 and –1 and, hence, no arbitrary transformation mechanisms are required in order to obtain symmetry with better than dead scores[5].

The utility scores derived for B5 and B10 allow us to test for proportionality and monotonicity in responses. If proportionality holds, and assuming no discounting then the utility value derived using B10 will be approximately equal to that derived using B5, that is, U(B10) ≅ U (B5). If utility is increasing or decreasing monotonically with duration, an increase in duration should increase utility in better than dead states and decrease utility in worse than dead states. For states ranked better than death a failure of montonicity is associated with MET, since this implies that utility increases initially with duration but after a point decreases.

Set two

After completing their ranking of the 10 profiles that make up set two, respondents were then asked to consider BBDD (BD10) in relation to the 'A' cards. For example, suppose that a respondent ranked BD10 below NDDD (A15), but above DDDD (A20), they would then be given cards A16 to A19 and asked to 'slot' these into their sub-ranking of these cards. This procedure was then repeated for BD5, again allowing for a test of proportionality. Up to this point, the procedure is exactly equivalent to that used in set one.

In set two however, as in conventional TTO approaches, we cannot derive worse than dead scores directly and we need to deploy some alternative 'worse than dead' procedure. To see why this is the case, consider figure 3; if 10 years blue is rated as worse than dead then BD10 will be ranked below A20, and we no longer have a mechanism for deriving scores. Likewise for all other durations of blue. Hence, we used a procedure equivalent to the 'worse than dead' method in the measurement and valuation of health study (MVH) used to derive TTO tariff values. Respondents who had *any* card from BD5 to B20 placed below A20 - dead for 20 years - were given a

[5] Of course changing the duration of the blue state in question would alter this. For example, setting B5 against A20, yields a lower bound of –3.

set of 'worse than dead cards' that had varying durations of the blue state *followed by* periods of normal health[6].

Of course, as in set one, it is possible that we obtain both better than, and worse that dead scores for the blue state, depending on the duration. This brings us back to the issue of MET. If there is no MET present in set two, and B5 is ranked higher than A20, then we would also expect that; B10 >A20, B5 >A20 and B20 > A20. In set two, the following pattern in the ranking would indicate that MET had occurred between 5 and 10 years in the blue health state; BD5 > A20, but that A20 > (BD10, BD15, BD20). In this case, we would derive a better than dead value for BD5 - five years in blue followed by death - but also have a worse than dead score for durations of 10 years and over.[7]

The representation used in set two allows us to test for the type of 'preference reversals' discussed in the literature on MET preferences. In set two, if BD5 > BD10, then we would also expect BD5 to be valued higher than BD10 (in relation to the 'A'-normal health and death - cards) and vice versa. As in set one, the utility scores derived for BD5 and BD10 allow us to test for proportionality and monotonicity in responses.

After completing both sets one and two, respondents were asked questions that allowed three separate estimates of their time preference. Each question consisted of a choice between 1 year in blue in x years time or 1 year of blue in y years time, where (x, y) took the values of (2, 6), (12, 16) and (2, 16) in each of the three questions. If respondents preferred to delay ill-health until period y, they were asked to increase the time spent in blue in y years until they were indifferent between the two choices. If they preferred instead to experience ill-health in period x, they were asked to increase the time spent in blue in x years until they were indifferent between the two choices.

---

[6] One such card BBNN (WD10) already appeared in the ranking and the position of WD10 relative to A20 dictated which cards the respondent saw next.

[7] Unlike with worse than dead scores in set one, the worse than dead scores here are not specific to particular durations as the worse than dead cards used here are 'generic'.

Study two results

The sample comprised of 41 respondents, 21 male, 20 female with a mean age of 21. All students were registered at Queen Mary University of London and were contacted through the Economics Department mailing list.

A discount rate was calculated for each respondent for each of the three discounting questions. The mean discount rate for the three questions was between 0.01 to 0.02 and the median was 0. These discount rates were used to adjust respondents' values.

Set one

Recall that the set one rankings show whether a particular duration of the blue health state is better than death (i.e. B5 > A5, B10 > A10 etc) or worse than death (i.e. B5 < A5, B10 < A10 etc). Recall also that MET arise in the ranking if shorter durations in blue are better than dead but longer durations of blue are worse than death (i.e. B5>A5, but B10<A10). In fact, the results show little evidence of such 'switches', the majority of respondents rating all durations in blue as either all better than, or all worse than dead (table 6). Table 6 shows that MET arose in 4/41 respondents. Some respondents also ranked shorter durations in blue as worse than death and longer durations as better than death- the opposite pattern to that predicted by MET. These 4 respondents may have found little benefit from 5 years of blue when it came at the end of the period, but preferred the 20 years in blue if the alternative was immediate death. A further 4 respondents have patterns within the ranking that were both better and worse than dead with no clear pattern (for example; B5<A5, B10>A10, B15>A15, B20<A20).

Table 6. Ranking of profiles- set one

	Respondents
All durations (B5-B20) better than dead	16
All durations (B5-B20) worse than dead	13
Shorter durations better than, longer durations worse than dead (i.e. MET)	4
Shorter durations worse than, longer durations better than dead (i.e. opposite to MET)	4
No clear pattern	4
Total	41

The subsequent sub-ranking's allows an estimate the utility value for 1 year of blue based on the 10 year and 5 year durations, B10 and B5 respectively. The summary statistics for these values are shown in table 7 with the median value for 1 year being 0.18 based on B10 and 0.15 based on B5 (to 2 decimal places). A two-tailed Kolmogorov-Smirnov test was used to test the hypothesis that the data was normally distributed. The test statistic was sufficiently large to reject the null hypothesis.

Table 7. Summary statistics for set one

	10 years blue U(B10)	5 years blue U(B5)	Discounted 10 years blue U(B10)	Discounted 5 years blue U(B5)
Observations	38*	40**	37*†	39**†
Mean	0.141	-0.145	0.156 to 0.181	-0.036 to -0.065
Median	0.175	0.150	0.206 to 0.213	0.212 to 0.217
Mode	-0.250	0.300		
Standard error	0.081	0.167	0.080 to 0.135	0.131 to 0.144
Standard deviation	0.497	1.055	0.480 to 0.519	0.843 to 0.899
25th percentile	-0.250	-0.575	-0.252 to -0.277	-0.511
75th percentile	0.575	0.500	0.613 to 0.621	0.531 to 0.568

* 3 respondents ranked B10- NNBB- below A20 and, hence, no score could be derived .
** 1 respondent did not answer this question
† 1 respondent did not provide a discount rate

We were also able to test the extent to which the values of B10 were proportional to the values of B5 -i.e. U (B10) = U(B5). Proportionality held for respondents who ranked both B10 and B5 as better than death using a 10% significance level.

Recall that a further test of MET arises from testing the extent to which utility is increasing or decreasing monotonically with duration. Table 8 shows that 14/37 respondents (5+8+1= 14) violate monotonicity in the direction that is consistent with MET. A further 3/37 respondents violate monotonicity with patterns that are opposite to that predicted by MET.

Table 8. Violations of monotonicity

	Respondents
Profiles that are better than death and where U(10 years blue) < = U(5 years blue)	5
Profiles that are worse than death and where U(10 years blue) > = U(5 years blue)	8
Profiles where U(5 years blue) is better than death but U(10 years blue) is worse than death	1
Profiles where U(5 years blue) is worse than death but U(10 years blue) is better than death	3
Total	17

Set two

Again the rankings show whether a profile (i.e. BD10) is better than death (ranked higher than A20) or worse than death (ranked lower than the A20 card). The majority of states are either all better than, or worse than dead, (table 9). In set two, MET arises in 10/41 respondents in the ranking exercise.

Table 9. Ranking of states- set two

	Respondents
All durations (BD5-B20) better than dead	20
All durations (BD5-B20) worse than dead	11
Shorter durations better than, longer durations worse than, dead (i.e. MET)	10
No clear pattern.	0
Total	41

The subsequent 'sub-ranking' procedures again estimates the value of 1 year of blue based on the 10 year and 5 year durations, BD10 and BD5 respectively. The summary statistics for these values are shown in table 10 and median value for 1 year was 0.25 and 0.30 based on BD10 and BD5 respectively. These preliminary summary statistics are based on unadjusted figures and include states worse than death whose scale could range from 0 to -39. A two-tailed Kolmogorov-Smirnov test was used to test the hypothesis that the data was normally distributed. The test statistic was sufficiently large to reject the null hypothesis.

Table 10. Summary statistics for set two

	10 years blue U(BD10)	5 years blue U(BD5)	Discounted 10 years blue U(BD10)	Discounted 5 years blue U(BD5)
Observations	41	38	40†	37†
Mean	-0.334	-0.163	-0.322 to -0.569	-0.114 to -0.178
Median	0.250	0.300	0.285 to 0.324	0.321 to 0.337
Mode	0.250	0.500		
Standard error	0.345	0.209	0.372 to 0.569	0.214 to 0.248
Standard deviation	2.209	1.287	2.352 to 3.599	1.301 to 1.505
25th percentile	-0.290	-0.382	-0.082 to -0.234	-0.318 to -0.325
75th percentile	0.475	0.500	0.488 to 0.526	0.546 to 0.564

† 1 respondent did not provide a discount rate

Further consideration of the degree of MET arises from the second sub-ranking procedure. For this analysis we exclude all respondents who ranked both BD5 and BD10 as worse than death with the aim to make our results comparable with earlier work (Dolan and Stalmeier, 2003)[8]. A test of monotonicity showed that 7/27 respondents violated monotonicity in the direction that is consistent with MET. We also tested the extent to which preference reversals occurred. Preference reversals occurred when BD5 > BD10 in the ranking, but B10 is valued higher than BD5 in the sub-ranking procedure. Preference reversals occurred in a total of 5/27 respondents.

Finally, we tested the extent to which the values of BD10 were proportional to the values of BD5 (i.e. B10=B5) using Wilcoxon's matched pairs tests. Proportionality again held for respondents who ranked BD5 and BD10 as better than death.

DISCUSSION

Existing models of intertemporal choice normally assume that preferences satisfy the formal condition of independence, or separability, which states that the value of a

[8] Moreover, checking for monotonicity in states worse than death did not follow easily given that the duration of blue is changed in worse than death states in set two.

sequence of outcomes equals the sum of the values of its component parts. Treadwell (1998) provided a test of preferential independence, as defined by Keeney and Raiffa, when health varies over time but under conditions of certainty. Even tests that were specifically designed to be more sensitive to independence violations, independence was still satisfied in the majority of cases. He concluded that independence held regardless of the discount rate that is used. In study one, we set out to see whether independence holds under conditions of uncertainty, what K&R term 'utility independence'. Like Treadwell, we found that independence holds in the majority of cases examined here.

We did, however, find a significant difference in the case of one of the tests - namely test five - whereby respondents were significantly more likely to set p higher in part A than part B. At first glance, this finding was puzzling, particularly as no such effect had been uncovered in test one, which was identical other than the length of time spent in the blue health state for sure. One tentative explanation for such an explanation may be the existence of MET preferences. Recall that, in moving between tests 5A and 5B in study one the amount of time spent in the blue state for sure changed from 5 to 10 years whilst the move from 1A to 1B changed from 10 to 15 years. Now, if respondents felt that 5 or so years in the blue state for sure was tolerable, but that all durations over that threshold are equally 'intolerable', this may explain the differences in results between tests one and five.

Of course, the robustness of the results would have to be tested before drawing any firm conclusions. In particular, we would wish to randomise the order in which tests one and five are undertaken in order to rule out learning effects. In conjunction with the responses to the qualitative questions, however, this finding did suggest that a further investigation of MET preferences would be worthwhile.

In study two, we use a ranking procedure in order to examine the effect of duration, with particular attention to the existence of MET preferences. A ranking procedure was adopted as this is relatively straightforward for respondents to do and we were keen to see the effect of asking respondents to value all profiles simultaneously. Previous research has shown ranking procedures to eliminate pervasive anomalies

from preferences over lotteries (see the Bateman et al)[9]. Our results showed that, although the blue health state was commonly rated as worse than dead, there were relatively few respondents exhibited MET preferences, rating the state either better than or worse than dead, regardless of duration. Hence, we find very few cases of the type of preference reversals uncovered elsewhere.

These results are in stark contrast to those of other studies. Sutherland et al (1982) investigated the extent to which maximal endurable time preference arose in a sample of 20 professionals from the Ontario cancer institute. They found that increasing the duration in the more severe states led to a preference for death over continued time in the dysfunctional state. Stalmeier et al (2001) asked two types of questions: a) a Time Trade Off questions - 'matching' tasks and b) a direct choice between a short and long period in the same health state - a choice task. There were less instances of MET preferences in the Time Trade-Off (TTO) questions compared to the direct choice: 14% (24/176) compared to 58% (103/176). Dolan and Stalmeier (2003) argued that TTO underestimates the degree of MET preferences as it encouraged respondents to answer as a proportion of the time remaining. Proportionality has been found to generally hold in the TTO method (Pliskin et al 1980, Hall et al 1992, Bleichrodt and Johannesson 1997), but Sackett and Torrance (1978) reject constant proportionality.

Our data shows that proportionality approximately holds for those respondents who rate states as better than dead, but we have no reason to suppose that this is a 'heuristic' as it was not found in conjunction with MET. We may speculate about why our findings fail to find the preference reversals reported in previous studies and there is probably a lot to be learned from the psychology of choice literature here. Of course, it simply may be that a sample of economic students are more likely to answer 'consistently' than other respondents and this is clearly something that needs further attention. Alternatively, it may be that the ranking procedure draws the respondent's attention to all aspect of the decision simultaneously and diminishes the scope for shifting the focus of attention between different tasks. As our scenarios make it explicit the number of life years that are being sacrificed in each case, making it less

[9] Whilst eliminating the common ratio effect from preferences over money lotteries, the ranking procedure was found to be susceptible to context effects.

likely that they will misconstrue the question. We consider a major advantage of the scenarios used in set one is that it allows worse than dead scores to be assessed in the same manner as better than dead scores.

There is clearly a lot more work to be done on this issue and analysis of our data set is ongoing. We do, however, think we have offered a useful method of eliciting preferences for health states that change over time.

Acknowledgements.

We would like to thank participants at the workshop in Alicante in October 2003 where the outline of study two was initially presented. We would also like to thank Graham Loomes, Peep Stalmeier, and Jordan Louviere for comments on the design. Financial support from the Medical Research Council is gratefully acknowledged.

Reference List

Bateman, I., Day, B., Loomes, G., Orr, S and Sugden, R, Does a ranking procedure eliminate the usual violations of expected utility theory?, Paper presented at PEG workshop, LSE 2002.

Bleichrodt, H., (1995). QALYS & HYEs: Under what conditions are they equivalent? *Journal of Health Economics 14*, 17-37.

Bleichrodt, H., & Quiggin, J. (1997). Characterizing QALYs under a general rank dependent utility model. *Journal of Risk of Uncertainty 15*, 151-165.

Bleichrodt, H. , Johannesson, M., 1997. An experimental test of constant proportional tradeoff and utility independence, Medical Decision Making 17, 21-32.

Dolan, P. & Stalmeier, P., (2003) The validity of time trade-off values in calculating QALYs: constant proportional time trade-off versus the proportional heuristic. *Journal of Health Economics*, 1-14.

Gafni, A., (1995). Time in health: can we measure individuals' "pure time preference". *Medical Decision Making 15*, 31-37.

Hall, J. Gerard, K., Salkeld, G. & Richardson, J. (1992). A cost utility analysis of mammography screening in Australia. *Social Science and Medicine*, 34, 993-1004.

Kahneman, D., Fredrickson, B.L., Schreibner, C.A., & Redelmeier, D.A. (1993). When more pain is preferred to less: adding a better end. *Psychological Science 4*, 401-405.

Keeney, R.L., & Raiffa, H. (1976*). Decisions with multiple objectives, preferences and value tradeoffs.* London: Wiley.

Loewenstein, G., & Prelec, D. (1993). Preferences for sequences of outcomes. *Psychological Review 100*, 91-108.

Miyamoto, J.M., Eraker, S., 1988. A multiplicative model of utility of survival duration and health quality, *Journal of Experimental Psychology:* General, 117, 3-20.

Pliskin, J.S., Shepard, D.S. & Weinstein, M.C. (1980). Utility function for life years and health status. *Operations Research* 28, 206-223.

Richardson, J., Hall, J., & Salkeld, G. (1996). The measurement of utility in multiphase health states. *International Journal of Technology Assessment in Health Care 12*, 151-162.

Ross, W.T., & Simonson, I. (1991). Evaluating pairs of experiences: a preference for happy endings. *Journal of Behavioral Decision Making 4*, 273-282.

Sackett, D.L. & Torrance, G.W. 1978. The utility of different health states as perceived by the general public. *Journal of Chronic Disease,* 31, 697-705.

Spencer, A. (2003). A test of the QALY model when heath states varies over time, *Social Science and Medicine* 57, 1697-1706.

Stalmeier, M. (1996). Proportional heuristics in time and conjoint measurement, *Medical Decision Making 16*, 36-44.

Stalmeier, P.F.M., Wakker, P.P. & Bezembinder, T.G.G. (1997) Preference reversals: violations of unidimensional procedure invariance. *Journal of Experimental Psychology: Human Perception and performance,* 23, 1196-1205.

Stalmeier, P.F.M., Chapman, G.B., de Boor, A.G.E.M. & van Lanschot, J.J.B. (2001) A fallacy of the multiplicative QALY model for low-quality weights in students and patients judging hypothetical health states, *International Journal of Technology Assessment in Health Care*, 17, 488-496.

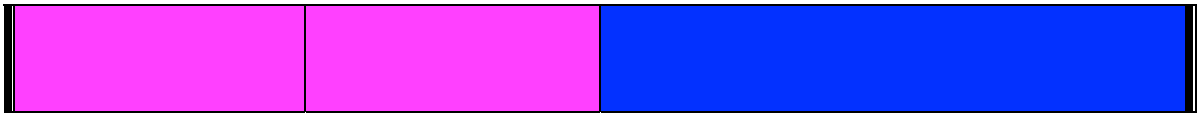
Sutherland, H.J., Llewellyn-Thomas, H., Boyd, N.F. & Till, J.E., (1982) Attitude toward quality of survival: the concept of maximal endurable time, *Medical Decision Making* 2, 299-309.

Treadwell, J.R., (1998). Tests of preferential independence in the QALY model. *Medical Decision Making 18*, 418-428.

Varey, C., & Kahneman, D. (1992). Experiences extended across time: evaluation of moments and episodes. *Journal of Behavioral Decision Making 5*, 169-185.

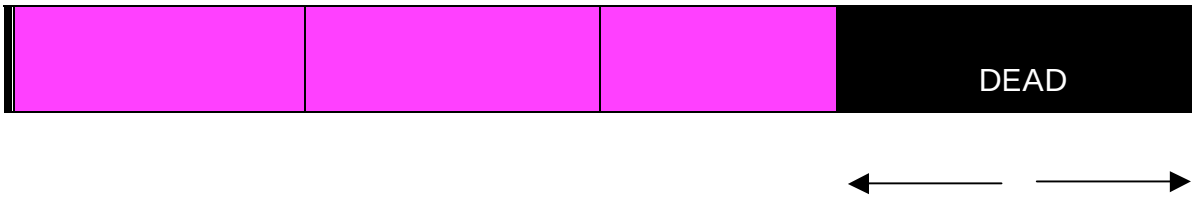
Figure 1: Derivation of better than dead scores in set one-B10

B10; In normal health until blue state for 10 years



Indifferent to;

AX; In normal health until death X years early.

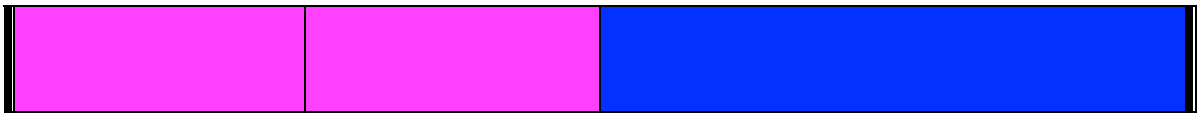


Then , 10+10B= (20-X), or, B= (10-X)/10.

When X<10, U(B) > 0.

Figure 2: Derivation of worse than dead scores in set one-B10

B10; In normal health until blue state for 10 years



Indifferent to;

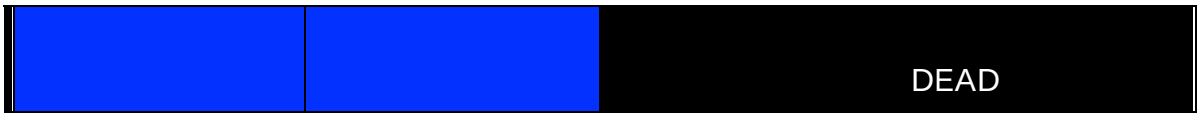
AX; In normal health until death X years early.

Again, 10+10B= (20-X), or, B= (10-X)/10.

When X<10, U(B) > 0. When X = 20, B = -1.

Figure 3: Derivation of better than dead scores in set two-B10

BD10; Blue state for 10 years, then death 10 years early.



Indifferent to;

AX; In normal health until death X years early.

Then , 10B = (20-X), or, B= (20-X)/10.

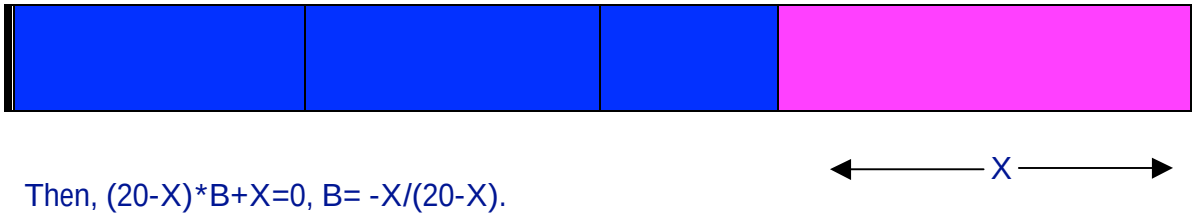
N.B. This is just a slightly different representation of a 'conventional' TTO question.

Figure 4: Derivation of worse than dead scores in set two

A20: In normal health until death 20 years early.

Indifferent to;

WDX; In blue state for 20-X years, then X years normal.



Then, (20-X)*B+X=0, B= -X/(20-X).
When X=10, B= -1. When X=19.5, B = -39.
N.B. This is equivalent to the worse than dead TTO scores used in the MVH study.