A methodological comparison of alternative standard gamble chaining procedures

Scotland GS, McNamee P

Health Economics Research Unit
University of Aberdeen
Polwarth Building
Foresterhill
Aberdeen
AB25 2ZD

Correspondence: g.scotland@abdn.ac.uk

Introduction

The standard gamble is the method for eliciting health state values implied from the axioms of von Neumann and Morgenstern's expected utility theory (EUT). This normative theory describes how rational individuals ought to make decisions when faced with uncertain outcomes. The method generally involves respondents being offered a hypothetical choice between a certain intermediate health outcome and an uncertain treatment option consisting of a chance of regaining full health but also a risk of immediate death. The aim is to elicit the probability of treatment success that renders the respondent indifferent between the certain and the uncertain option. This probability (p) of indifference represents the utility value that the respondent places on the intermediate health state on a scale bounded by death (0) and full health (1). Gambles of this type, where full health and death are used as the success and failure outcomes respectively, are often termed standard reference gambles. A problem of this method is that respondents can be unwilling to risk any chance of death when the health state being valued is relatively minor. This unwillingness to accept any chance of death leads to minor health states being valued equal to full health, despite respondents indicating that they feel these states are worse than full health in ranking and rating exercises.

A variation of the standard gamble that is commonly used to overcome this insensitivity, involves splitting the process into two or more separate questions. First of all a minor health state is valued relative to full health and a more severe health state (not death). Secondly, the more severe health state is valued relative to full health and death, allowing the minor health state to be linked indirectly to the death to full health scale. This process is known as chaining through the failure outcome. Values obtained in this way are sometimes referred to as indirect, as opposed to direct values obtained from reference gambles.

Though generally only used for assessing minor and temporary states, there are times when it may also be desirable to use chaining to value severe health states. For example, it can help make the process seem more realistic to the respondent. However, the independence axiom of expected utility theory predicts that preferences for the same intermediate health state should be the same whether valued directly through one single gamble or indirectly through chaining (procedural invariance). Therefore, the comparison of direct and indirect values for severe health states offers a way to test the internal consistency of the standard gamble. If it is internally consistent, then indirect values obtained by chaining through the failure outcome,

and direct values obtained from reference gambles, should not differ significantly or systematically from one another.

Unfortunately it has been demonstrated that there is a tendency for values elicited indirectly, by two-staged gambles chained through the failure outcome, to exceed those elicited directly through standard reference gambles (Llewellyn-Thomas et al., 1982; Bleichrodt, 2001; Chilton and Spencer, 2001; Spencer, 2001; Stalmeier, 2002; Oliver 2003, 2004). This is a violation of the independence axiom of expected utility theory.

Prospect theory has been proposed as an alternative theory of behaviour under uncertainty, which can accommodate such violations of EUT (Kahneman and Tversky, 1979; Tversky and Kahneman 1986, 1992). It proposes two major modifications to EUT. First of all, it stipulates that the carriers of value are gains and losses as opposed to final end points. Applied to the standard gamble, it is hypothesised that individuals view the success and failure outcomes of the risky treatment option as potential gains and losses from the reference point of the certain treatment outcome. It has also been noted that individuals are often loss averse – i.e. the disutility they suffer from losses is of greater magnitude than the utility they experience from gains of the exact same size. In addition, prospect theory suggests individuals become less sensitive to gains and losses as these increase from the reference point (diminishing marginal sensitivity).

The second major modification proposed by prospect theory is that individuals transform probabilities when faced with risky options that involve losses and gains. The observed pattern is that individuals tend to overweight small probabilities and underweight large probabilities.

Oliver (2003) tested the consistency of the standard gamble while adjusting for these modifications proposed by prospect theory. He did this by incorporating previously estimated weights for loss aversion and probability transformation into the standard gamble valuation equation. The finding was that the incorporation of loss weighting alone offered the only improvement to internal consistency. Oliver (2004) explains how the principal of loss aversion may result in individuals placing particular emphasis on the failure outcome in the standard gamble. As a result, he argues, the failure outcome may always be viewed as a substantial loss and, due to the principal of diminishing marginal sensitivity, the change in disutility when one substantial loss

is replaced with another, will be insufficient to ensure internal consistency - i.e. the failure outcome in standard reference and chained to failure gambles will be viewed as being too similar to ensure internal consistency.

The hypothesised mechanism underlying internal inconsistency described by Oliver (2004), suggests that internal consistency might be improved if the extent to which loss aversion and diminishing marginal sensitivity impact upon respondents reasoning process is reduced. Although loss aversion is often considered to be a constant, there is some evidence to suggest that it may vary across different settings and trade-offs (Bleichdort and pinto, 2002). This suggests that it might be possible to chain through the failure outcome in a way that minimises the impact of loss aversion and thus improves the internal consistency of the standard gamble. We hypothesised that the impact of loss aversion and diminishing marginal sensitivity may vary depending on the outcomes and number of steps through which the chaining occurs. For a set of health states being valued, values can be chained through the health state considered worst in the set. Alternatively, chaining can occur through the health state considered next best to the one being valued. Finally, with the latter approach, chaining can occur in a two-stage process, or a multistage process depending on the number of health states being valued. For example, if there four chronic health states in the set, where H_1 H_2 H_3 H_4 (means preferred to) then H_1 could be valued relative to H_2 and then chained back to the zero to one scale through a gamble where H_2 is valued relative to full health and death. Alternatively, H_2 could be valued relative to full health and H_3, then H_3 valued relative to full health and H_4, and finally H_4 valued relative to full health and death. The final gamble provides the link back to the zero to one scale for all the previous gambles.

The purpose of the study was to assess whether the above approaches for chaining through the failure outcome influence the internal consistency of the standard gamble. The study was carried out in the context of a valuation exercise to elicit women's preferences for adverse birth outcomes associated with twin pregnancy.

Methods

Eighty-one women attending the Assisted Reproduction Unit (Aberdeen Maternity Hospital) experiencing fertility problems were invited to participate in a valuation interview. All women waiting to undergo intra uterine insemination or in-vitro fertilisation procedures were considered eligible for inclusion. Four birth outcomes were chosen for valuation on the basis of having a higher chance of occurring with

twin pregnancy, a risk associated with assisted reproductive techniques. These were perinatal death (PD), giving birth to a child with cerebral palsy (CP), giving birth to a child with cognitive impairments (CI), and giving birth to a child with visual impairments (VI). Descriptions were developed for each of the outcomes through literature review and consultation with clinical experts (Furlong et al, 1990), and labelling of states was avoided (Froberg and Kane, 1989). Descriptions for the three child disability outcomes were based on the descriptive system of the Health Utilities Index Mark 2 (HUI 2) instrument, which was originally designed to value the health of survivors of childhood cancer (Feeny et al, 1992). Perinatal death was framed as a chronic outcome in that it was described as being followed by no subsequent pregnancy for the rest of the woman's life. Women were asked to view all the health state descriptions from their own perspective – and to imagine how experiencing each of the adverse outcomes would affect their own quality of life. Since it is potential mothers who have to make the choices regarding treatment, we wanted to understand the relative values women placed on experiencing each of the different birth outcomes. A lifetime perspective was taken for all valuations.

Prior to the standard gamble exercise, participants were asked to rate each of the birth outcomes on a visual analogue scale. Two additional states were added to act as reference states for the scale. Full health for mother and baby was used as the top anchor (100) and immediate death of mother and unborn baby as the bottom anchor (0). The main purpose of the VAS exercise was to get subjects thinking about their strengths of preference for the outcomes before commencing the standard gamble exercise. The results were used to rank the four birth outcomes (CP, CI, VI, and PD) from H_1 (most preferred) to H_4 (least preferred).

Standard Gamble questions

Women were asked to imagine that they had become pregnant but had developed some complications. They were offered a choice between receiving a treatment (alternative 1) or leaving the complications untreated (alternative 2). They were told to imagine that if they were to choose no treatment (alternative 2) the pregnancy would definitely result in one of the adverse birth outcomes. Alternative 1 (the treatment), on the other hand, was presented as having a chance (p) of success (resulting in full health for mother and baby) but also a chance (1-p) of failure (resulting in immediate death for the mother and unborn baby). The probability (p) that rendered the individual indifferent between the certain and uncertain options was elicited in one of two ways: either by systematically varying the probability until the

respondent became indifferent between the two options (format 1), or by asking subjects to state the minimum probability of achieving the success outcome that they would be willing to accept in order to choose the uncertain option over the certain one (format 2). Women were randomly assigned to one or other of the elicitation formats. For the probability varying method, the interview was supplemented using a chance board (Furlong et al, 1990).

The order of presentation of the standard gamble questions is described in Table 1. First of all, the four chronic birth outcomes (CP, CI, VI, and PD) were valued relative to giving birth to a healthy child (HC) and immediate death (D). Following this, the three highest ranked outcomes (H_1 to H_3) were valued relative to giving birth to a healthy child and H_4. Finally, H_1 was valued relative to giving birth to a healthy child and H_2, and H_2 valued relative to giving birth to a health child and H_3.

These combinations give rise to three different chained procedures, as described in Table 2. First, H_1 to H_3 were chained through a 2-link procedure using H_4, the worst ranked outcome (2-link^W). Second, H_1 and H_2 were chained in a 2-link procedure via the next best outcome, H_2 and H_3 respectively (2-link^B). Finally, a multi-link chaining procedure was tested where values for all the states were solved in sequence from worst to best – each time using the value obtained from the previous gamble to link the more preferred state to the zero to one scale (Multi-link).

The different procedures to valuing the outcomes can be solved using the standard gamble equation U(H) = pU(HC)+(1-p)U(D), where U(H) is the value of the health outcome being valued, U(HC) is the value associated with giving birth to a healthy child, U(D) is the value associated with immediate death, and p is the probability that renders the respondent indifferent between the certain and uncertain options.

For the standard reference gambles - e.g. Direct $U(H_2)$:

$$U(H_2) = p_2(HC)+(1- p_2)U(D)$$

If U(HC) = 1 and U(D) = 0, this expression becomes:

$$U(H_2) = p_2$$

For values chained through the worst ranked outcome – e.g. 2-link ^{W} $U(H_2)$

$U(H_2) = p_6(HC)+(1- p_6)U(H_4)$

If $U(HC) = 1$ and $U(H_4) = p4$ from question 4, this expression becomes:

$U(H_2) = p_6+(1- p_6)\, p_4$

For values chained through the next best outcome – e.g. 2-link ^{B} $U(H_2)$

$U(H_2) = p_9(HC)+(1- p_9)U(H_3)$

If $U(HC) = 1$ and $U(H3) = p3$ from question 3, this expression becomes:

$U(H_2) = p_9+(1- p_9)\, p_3$

For values chained using the multilink approach – e.g. Multilink $U(H_2)$

$U(H_2) = p_9(HC)+(1- p_9)$ 2-link ^{W} $U(H_3)$

If $U(HC) = 1$ and 2-link ^{W} $U(H_3) = p_7+(1- p_7)\, p_4$ from questions 7 and 4, this expression becomes:

$U(H_2) = p_9+(1- p_9)\,(p_7+(1- p_7)\, p_4)$

Data analysis

Tests of internal consistency were conducted by assessing whether direct values for the health outcomes were significantly different from indirect values, and whether indirect values were significantly different from each other.

In other words, the following hypotheses were tested:

1. Null hypothesis: Direct $U(H_i) =$ {2-link ^{W} $U(H_i)$, 2-link ^{B} $U(H_i)$, Multilink $U(H_i)$};
Alternative hypothesis: Direct $U(H_i) \neq$ {2-link ^{W} $U(H_i)$, 2-link ^{B} $U(H_i)$, Multilink $U(H_i)$};

2. Null hypothesis: {2-link ^{W} $U(H_i)$, 2-link ^{B} $U(H_i)$, Multilink $U(H_i)$} = {2-link ^{W} $U(H_i)$, 2-link ^{B} $U(H_i)$, Multilink $U(H_i)$};
Alternative hypothesis: {2-link ^{W} $U(H_i)$, 2-link ^{B} $U(H_i)$, Multilink $U(H_i)$} ≠ {2-link ^{W} $U(H_i)$, 2-link ^{B} $U(H_i)$, Multilink $U(H_i)$}

Data were analysed using SPSS version 13. Wilcoxen signed rank tests were used to compare reference values and indirect values for the health states obtained through the different chaining procedures.

Results

Of 226 women invited to participate in the valuation interview, 81 agreed to take part. The mean age of the women was 32 years (SD 4.1), 93% had no existing children, and 84% were educated beyond the minimum school leaving age. Seven interviews were dropped from the analysis because the rank ordering of preferences obtained from the rating scale exercise was inconsistent with the rank ordering implied from the standard gamble reference values.

The results of the tests of consistency for each different format of the interview are presented in Tables 3 and 4. The results indicate that all methods of chaining through the failure outcome result in values significantly higher than values elicited directly from reference gambles. The results also indicate the multilevel chaining procedures produce values significantly higher than two-stage chaining procedures. The comparison of the two-stage chains linked through worst ranked and next best outcomes appears inconclusive. Using the standard approach to elicit the probability of indifference (format 1), the values from two-link chains using the next best outcome (2-link ^{B}) are significantly higher than two-link chains using the worst ranked outcome (2-link ^{W}). When format 2 was used, there were no significant differences in the values obtained using these two procedures, though there was a tendency for values from two-link chains using the next best outcome (2-link ^{B}) to be lower.

Discussion

The results suggest that values chained through the failure outcome are systematically higher than standard reference values. This is consistent with previous findings (Llewellyn-Thomas et al., 1982; Bleichrodt, 2001; Chilton and Spencer, 2001; Spencer, 2001; Stalmeier, 2002; Oliver 2003, 2004). We also found that chaining procedures using multiple links produce significantly higher values than 2-stage chaining procedures. This is also consistent with previous findings (Morrison, 2000).

The main explanation offered in the literature over why values chained through the failure outcome are higher than standard reference gambles relate to loss aversion and diminishing marginal sensitivity. The presence of both factors prevents respondents accepting a sufficiently higher level of risk when the failure outcome is switched to a less severe outcome during the chaining procedure. It is hypothesised

that respondents view the success and failure outcomes of gambles as gains or losses from the reference point of the certain outcome. The principal of 'loss aversion' implies that the disutility individuals suffer from losses is of greater magnitude than the utility they experience from gains of the exact same size. The presence of this effect may lead individuals to focus on the loss rather than the gain. Coupled with diminishing marginal sensitivity (the observation that the marginal perceived impact of a change in outcome diminishes with the size of the outcome), respondents may overweight a movement from the certain health state to more minor states, relative to a movement to a more severe state. This explanation is consistent with our finding that chained values are higher than reference values and it can also explain why multilink chains produce higher values than 2-link chains. The multilink chains would be expected to produce higher values if loss aversion and diminishing marginal sensitivity impact upon each step in the chaining procedure.

The comparison of values from the different two-stage approaches to chaining (2-link^{W} and 2-link^{B}) produced mixed results. When using the classical approach to elicit the probability of indifference, values chained through the next best outcome were significantly higher than values chained through the worst ranked outcome (Table 3). Although this finding may be explained by loss aversion and diminishing marginal sensitivity, an alternative reason could relate to question order. An ordering effect may have been induced as the questions that involved chaining through the next best outcome were always carried out last. It is possible that the previous questions, which involved switching of the failure outcome, led to greater focus on the failure outcome as the interview progressed. This might have led individuals to use a decision heuristic based on previous response patterns that would have the same effect as loss aversion. One solution would have been to randomise the question order, but this may have made the interviews more difficult for respondents to complete.

On the other hand, chaining through the next best outcome may be expected to produce a systematically higher degree of inconsistency, regardless of ordering, compared to chaining through the worse outcome, through the presence of loss aversion and diminishing marginal sensitivity.

To illustrate this, if we first value three health outcomes (H_1 to H_3) relative to full health and death, then the direct reference values for the three states are given by the equation $U(H_i) = pU(FH) + (1-p)U(D)$. If full health and death are normalised to

one and zero respectively, this simplifies to $U(H_i) = p$. Alternatively, H_1 can be valued on the zero to one scale indirectly using a gamble where the outcomes of the risky treatment option are full health and H_3 (2-link^W), or full health and H_2 (2-link^B). We can then estimate the indirect values by substituting the reference values for H_2 or H_3 into the equation $U(H) = pU(FH) + (1-p)U(H_2 \text{ or } H_3)$. Suppose an individual values H_1, H_2 and H_3 at 0.8, 0.6 and 0.4 respectively in standard reference gambles. In order for their indirect value for H_1 (chained through H_3) to be consistent with their reference value, they would have to decrease their probability of indifference in the new gamble to 0.666. Any amount higher than 0.66 would then produce an indirect value higher than 0.8. However, because individuals are loss averse, and insensitive to changes in the loss outcome, they may be unwilling to decease their probability of indifference to this level. Now suppose the same individual values H_1 relative to full health and H_2. In order for their indirect value for H_1 to be equal to the direct value of 0.8, they would now have to reduce their probability of indifference to 0.5. Our findings suggest that such adjustment of the probability of indifference in this new gamble is even more inadequate. This might be because, as diminishing marginal sensitivity predicts, the smaller the loss, the more individuals are likely to overweight it. Such overweighting essentially makes H_2 appear closer to H_3 than it ought to be, leading to a larger upward bias in the probability of indifference for the new gamble.

To our knowledge, this is the first time that indirect values obtained by chaining through failure outcomes of varying severity have been compared. As mentioned above, we cannot say conclusively whether the differences are induced by loss aversion and diminishing marginal sensitivity, or due to an ordering effect. Adding to the uncertainty is the finding that there were no significant differences between the two-link chaining methods when a different questionnaire format was used to elicit the probability of indifference. This might be because, when using this format, chaining through the worst ranked outcome produced values that were too high to allow further inconsistency to be observed. Alternatively, this particular format may make it easier for individuals to adjust their probability of indifference and thus improve the consistency. Further research is required to assess whether the same pattern of results is observed when controlling for ordering effects. If these results were repeated, it would suggest that if chaining through the failure outcome is to be used, it should be conducted using the approach that minimises internal inconsistency – i.e. using a failure outcome that the respondent perceives to be the worst state, rather than only the next worst to the health state being valued.

References

Bleichrodt H. Probability weighting in choice under risk: An empirical test. Journal of risk and uncertainty 2001; 23: 185-198

Bleichrodt H, Pinto JL. Loss aversion and scale compatibility in two attribute trade-offs. Journal of Mathematical Psychology 2002; 46: 315-337

Chilton S, Spencer A. Empirical evidence of inconsistency in standard gamble choices under direct and indirect elicitation methods. Swiss Journal of Economics

Feeny D, Furlong W, Barr RD, Torrance GW, Rosenbaum P, Weitzman S. A coprehensive multiattribute system for classifying the health status of survivors of childhood cancer. J Clin Oncol 1992; 10:923-28

Froberg DG, Kane RL. Methodology for measuring health-state preferences - I: Measurement strategies. J Clin Epidemiol 1989; 42:345-54.

Furlong W, Feeny D, Torrance G, Barr R, Horsman J. Guide to design and development of health-state utility instrumentation. Hamilton, Canada: McMaster University, 1990. CHEPA Working Paper; 90.

Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. Economietrica1979; 47: 263-291

Llewellyn-Thomas H, Sutherland HJ, Tibshirani R, Ciampi A, Till JE, Boyd NF. The measurement of patients values in medicine. Medical Decision Making 1982; 2: 449-462

Morrison GC. The Endowment effect and expected utility. Scottish Journal of Political Economy 2000; 47: 183-197

Oliver A. The internal consistency of the standard gamble: tests after adjusting for prospect theory. Journal of Health Economics 2003; 22: 659-674

Oliver A. Testing the internal consistency of the standard gamble in ‘success’ and ‘failure’ frames. Soc Sci Med 2004; 58: 2219-2229

Spencer A. The implications of linking questions within the SG and TTO methods. Working Paper No. 438; Queen Mary University of London, 2001

Stalmeier PFM. Discrepancies between chained and classic utilities induced by anchoring with occasional adjustments. Medical Decision Making 2002; 22: 53-64

Tversky A, Kahneman D. Rational choice and the framing of decisions. Journal of Business 1986; 59: S251-278

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

Table 1: Standard gamble questions

Question	**Health state**	**Comparison**	**Probability of indifference**
1	H_1	FH + D	P_1
2	H_2	FH + D	P_2
3	H_3	FH + D	P_3
4	H_4	FH + D	P_4
5	H_1	FH + H4	P_5
6	H_2	FH + H4	P_6
7	H_3	FH + H4	P_7
8	H_1	FH + H2	P_8
9	H_2	FH + H3	P_9

Table 2: Questions used to test the different chining procedures

Health outcome	**Direct values**	**2-link ^{W} chains**	**2-link ^{B} chains**	**Multilink chains**
H1	Q1	Q5 & Q4	Q8 & Q2	Q8, Q9, Q7, & Q4
H2	Q2	Q6 & Q4	Q9 & Q3	Q9, Q7, & Q4
H3	Q3	Q7 & Q4		
H4	Q4			

Table 3: Reference versus two-link and two-link versus multi-link chains (Format 1)

Direct vs 2-link chains (worst)						
	Response pattern					
Outcome	Direct< 2-link ^{W}	Direct= 2-link ^{W}	Direct> 2-link ^{W}	Median Direct	Median 2-link ^{W}	Wilcoxen matched pairs
CP	24	1	6	0.875	0.904	P<0.01
CI	25	2	5	0.925	0.954	P<0.01
VI	24	2	6	0.925	0.964	P<0.01
2-link chains (worst) vs 2-link chains (next best)						
Outcome	2-link ^{W}< 2-link ^{B}	2-link ^{W}= 2-link ^{B}	2-link ^{W}> 2-link ^{B}	Median 2-link ^{W}	Median 2-link ^{B}	Wilcoxen matched pairs
CP	3	0	0	0.904	0.947	ns
CI	16	4	4	0.954	0.968	P<0.05
VI	18	4	5	0.964	0.978	P<0.01
2-link chains (worst) vs multilink chains						
Outcome	2-link ^{W} < multilink	2-link ^{W} = multilink	2-link ^{W} = multilink	Median 2-link ^{W}	Median multilink	Wilcoxen matched pairs
CP	3	0	0	0.904	0.954	ns
CI	23	0	0	0.954	0.982	P<0.001
VI	26	0	1	0.964	0.987	P<0.001

Table 4: Reference versus two-link and two-link versus multi-link chains (Format 2)

Classic vs 2-link chains (worst)						
	Response pattern					
Outcome	Direct< 2-link ^{W}	Direct= 2-link ^{W}	Direct> 2-link ^{W}	Median Direct	Median 2-link ^{W}	Wilcoxen matched pairs
CP	20	10	6	0.900	0.975	P<0.01
CI	15	16	5	0.97	0.997	P<0.05
VI	14	17	6	0.97	0.999	ns
2-link chains (worst) vs 2-link chains (next best)						
Outcome	2-link ^{W}< 2-link ^{B}	2-link ^{W}= 2-link ^{B}	2-link ^{W}> 2-link ^{B}	Median 2-link ^{W}	Median 2-link ^{B}	Wilcoxen matched pairs
CP	1	2	2	0.975	0.99	ns
CI	10	9	11	0.997	0.986	ns
VI	11	15	9	0.999	0.997	ns
2-link chains (worst) vs multilink chains						
Outcome	2-link ^{W} < multilink	2-link ^{W} = multilink	2-link ^{W} = multilink	Median 2-link ^{W}	Median multilink	Wilcoxen matched pairs
CP	3	2	0	0.975	1	ns
CI	16	11	3	0.997	0.997	P<0.01
VI	19	14	2	0.999	1	P<0.001