Adam J. Oliver


# Testing the internal consistency of the standard gamble in 'success' and 'failure' frames


**Originally published in Social science & medicine, 58 (11). pp. 2219-2229 © 2004 Elsevier.**

**TESTING THE INTERNAL CONSISTENCY OF THE STANDARD GAMBLE IN 'SUCCESS' AND 'FAILURE' FRAMES**

ADAM OLIVER*
LSE Health and Social Care, London, UK

* Correspondence to: LSE Health and Social Care, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK. Tel. +44 (0)20 7955 6471; Fax. +44 (0)20 7955 6803; E-mail: a.j.oliver@lse.ac.uk

**Abstract**

Decision making behaviour has often been shown to vary following changes in the way in which choice problems are described (or ‘framed’). Moreover, a number of researchers have demonstrated that the standard gamble is prone to internal inconsistency, and loss aversion has been proposed as an explanation for this observed bias. This study attempts to alter the influence of loss aversion by framing the treatment arm of the standard gamble in terms of success (where we may expect the influence of loss aversion to be relatively weak) and in terms of failure (where we may expect the influence of loss aversion to be relatively strong). The objectives of the study are (1) to test whether standard gamble values vary when structurally identical gambles are differentially framed, and (2) to test whether the standard gamble is equally prone to internal inconsistency across the two frames (i.e. across frames where we may expect the strength of loss aversion to differ). The results show that compared to framing in terms of treatment success, significantly higher values were inferred when the gamble was framed in terms of treatment failure. However, there was no difference in the quite marked levels of internal inconsistency observed in both frames. It is possible that the essential construct of the standard gamble induces substantial loss aversion irrespective of the way in which the gamble is framed, which offers a fundamental challenge to the usefulness of this value elicitation instrument. It is therefore recommended that further tests are undertaken on more sophisticated corrective procedures designed to limit the influence of loss aversion.

*Keywords:* Standard gamble; Internal consistency; Framing; Expected utility theory

## Introduction

It has been argued that the standard gamble is the gold standard for eliciting health state values (Torrance, 1986; Torrance & Feeny, 1989). The reason for this is that health care decisions invariably involve a degree of risk, and unlike other often used elicitation methods (e.g. the time trade-off and the rating scale), the standard gamble is implied from the assumptions underlying expected utility theory (EU). The standard gamble thus has a firm grounding in the theory of risk and uncertainty.

The standard gamble generally requires a respondent to trade off the certainty of being in an intermediate health state for her remaining life expectancy with a 'treatment' that offers a chance of regaining full health for her remaining life expectancy but that also entails a risk of immediate death. Note that standard gambles that use full health and death as the success and failure outcomes, respectively, are sometimes termed 'basic reference' gambles (Rutten-van Molken et al., 1995; Spencer, 2001). The respondent is asked to state the probability of treatment success that she would require in order for her to be indifferent between remaining in the intermediate health state and accepting the treatment. The stated probability then represents the value that the respondent places upon the intermediate health state.

When using immediate death as the treatment failure outcome, a possible problem with the standard gamble is that many people may not be willing to accept any chance of treatment failure when minor or temporary states of poor health are being valued. For example, if a respondent has (or is being asked to assume that she has) a slight limp, she may not be willing to accept a treatment that entails any chance of immediate death. In these circumstances, the implication is that minor and temporary health states may often be valued as highly as full health, simply because the basic reference standard gamble is insufficiently sensitive to capture true underlying preferences. Indeed, in valuing minor health states, it has been empirically demonstrated that respondents will often be unwilling to trade chances of survival for improvements in health status, even though the same respondents rated those minor health states lower than full health in a simple rating exercise (Jones-Lee, Loomes, & Philips, 1995).

One possible way of overcoming the problem of insufficient sensitivity in the standard gamble is by indirectly linking - or 'chaining' (Rutten-van Molken et al., 1995; Spencer, 2001) - minor or temporary health states to death (Torrance, 1986; Jones-Lee et al., 1993). Thus, when valuing a minor or temporary health state, a non-fatal health outcome (that is more severe than the intermediate health state that is being valued) could be used instead of immediate death as the treatment failure outcome. For example, if a slight limp is the health state being valued, the treatment failure outcome could be the loss of a leg. The value of the loss of a leg could then be chained in from a further gamble where the loss of a leg is valued against a treatment that offers a chance of full health or immediate death.

Moderate and severe health states (as opposed to minor and temporary health states) are usually valued against full health and death, basically because it is generally assumed that the standard gamble is sufficiently sensitive for these purposes (Torrance, 1986). However, in order to test whether the chaining process is unbiased (and thus appropriate for valuing minor and temporary health states), moderate and severe health states can be valued both directly against full health and death, and through a chaining exercise. If the standard gamble is internally consistent (i.e. if chaining appears to give unbiased answers), then, for any particular moderate or severe health state, direct value elicitation through a basic reference exercise and indirect value elicitation through a chained exercise should generate values that do not significantly or systematically differ from one another.

## Background

Unfortunately, the standard gamble has been empirically observed to be internally inconsistent (Llewellyn-Thomas, Sutherland, Tibshirani, Ciampi, Till, & Boyd, 1982; Rutten-van Molken, Bakker, van Doorslaer, & van der Linden, 1995; Bleichrodt, 1996; 2001; Morrison, 1996; Spencer, 2001). To understand the concept of internal consistency a little more formally, consider Figure 1.

[Insert Figure 1]

In Figure 1, (a), (b) and (c) are all standard gambles. In (a), assume that an individual is given a choice between remaining in a 'higher intermediate' health state for her remaining life expectancy, and a treatment that offers an unknown probability, p, of regaining full health for her remaining life expectancy otherwise immediate death. Therefore, v(higher intermediate) = pv(full health) + (1-p)v(death), where v(.) is the standard gamble value function. If the values of full health and death (i.e. the best and worst outcomes) are respectively normalised at one and zero, then v(higher intermediate) = p.

In (b), a similar standard gamble design is presented, but the 'higher intermediate' health state has been replaced by a worse health state, termed the 'lower intermediate' health state. The individual's stated probability of indifference is now given by q. Following the same process as described for (a), v(lower intermediate) = q.

Finally, in (c), the individual is asked to choose the probability, r, for which she would be indifferent between the certainty of remaining in the higher intermediate health state for her remaining life expectancy, and a treatment that offers her the chance of regaining full health for her remaining life expectancy otherwise suffering the lower intermediate health state for her remaining life expectancy. In this case, v(higher intermediate) = rv(full health) + (1-r)v(lower intermediate). If the standard gamble is internally consistent, we ought to be able to substitute into this equation the previously elicited values for the higher and lower intermediate health states. Hence, the individual should state an r commensurate with p = r + (1-r)q; i.e. r = (p-q)/(1-q).

The process outlined above chains to the failure outcome, implying that the exercise involves replacing the failure outcome in the gamble with a health state that is better than the original failure outcome. Alternatively, it is possible to chain to the success outcome, which would mean that an exercise is undertaken where the original success outcome is replaced by an inferior health state. The studies that have tested the internal consistency of the standard gamble in the context of health have generally found that although the standard gamble is internally consistent when chaining to the success outcome, there is a systematic tendency for (indirect) chained to failure values to exceed (direct) basic reference values (Llewellyn-Thomas et al., 1982; Rutten-van Molken et al., 1995; Bleichrodt, 1996; 2001; Morrison, 1996; Spencer, 2001); i.e. in the context of the hypothetical example given above, for r + (1-r)q to exceed p.

A number of explanations have been suggested for the observed internal inconsistency in the standard gamble. For example, Bleichrodt (1996) tested for (i) the possible influence of a change in the most salient outcome, (ii) imprecise preferences and (iii) probability weighting when chaining to the failure outcome in the standard gamble. The hypothesis of a change in the most salient outcome is based on the assumption that people focus on the successful treatment outcome in the basic reference gambles, but then switch their focus to the failure outcome when this outcome is altered during the process of chaining. Bleichrodt hypothesised that the change in focus from the success outcome to the failure outcome causes a greater level of risk aversion in gambles that are chained to death, and attempted to control for this possible effect by encouraging his respondents to focus on the success and failure outcomes in all gambles. He continued to observe significant levels of internal inconsistency after controlling for this hypothesised bias.

The imprecise preferences explanation is based on the fact that people are faced with a task (i.e. the standard gamble) that is unfamiliar to them. Bleichrodt controlled for this potential bias by allowing his respondents to offer a range of indifference probabilities, and constructed a personal confidence interval for each individual around their range. When an individual's personal confidence intervals for a basic reference value and a chained value for a particular health outcome did not overlap, Bleichrodt concluded that the observed internal inconsistency could not be explained by imprecise preferences. He found that a large proportion of his respondents did not have overlapping confidence intervals, and for those who did not, the direction of the difference between the basic reference values and the chained values continued to be systematic (i.e. the values chained to death tended to exceed the basic reference values). The observed internal inconsistency could not, therefore, be explained by imprecise preferences.

Finally, the probability weighting explanation is based on the assumption that people weight (or 'transform') 'raw' probabilities. Bleichrodt applied the commonly observed pattern of an inverse S-shaped probability weighting function (i.e. that people overweight small probabilities and underweight large probabilities) to his respondents' responses (for evidence of an inverse S-shaped probability weighting function, see, for example, Tversky & Kahneman, 1992; Camerer & Ho, 1994; Wu & Gonzalez, 1996; Bleichrodt, van Rijn & Johannesson, 1999). Unfortunately, this application actually caused even greater internal inconsistency in the observed responses, leading Bleichrodt to imply in a later article that the inconsistency is more likely to be explained by 'loss aversion' (Bleichrodt, 2001).

Loss aversion has been extensively observed and noted in the literature (e.g., Fishburn & Kochenberger, 1979; Kahneman & Tversky, 1979; 1984; Hershey & Schoemaker, 1980; Payne, Laughhunn, & Crum, 1980; Eraker & Sox, 1981; Tversky & Kahneman, 1981; 1991; 1992; Fischhoff, 1983; Enemark, 1994), and implies that the disutility that individuals suffer from losses is of significantly greater magnitude than the utility they enjoy from gains of the same absolute size. In other words, loss aversion means that losses loom larger than gains. This psychological behavioural pattern is not accounted for by EU or, by implication, the standard gamble, but if people perceive the success outcome in the treatment arm of the standard gamble as a gain and the failure outcome as a loss, loss aversion may have a significant influence on the values elicited with this instrument. Bleichrodt et al. (2001) assume that the 'reference point', from which people respectively perceive the success and failure outcomes as gains and losses, is the certain, intermediate health outcome in the standard gamble. Assume that the success outcome in the gamble is given by $x_1$, the failure outcome by $x_3$, and the certain intermediate outcome by $x_2$. Simplifying from Bleichrodt et al. (2001), the standard gamble valuation formula with loss aversion is given by:

$$v(x_2) = v(x_2) + p(v(x_1)-v(x_2)) - \lambda(1-p)(v(x_2)-v(x_3)) \quad (1)$$

where $\lambda$, the 'loss weighting parameter', is the factor by which people weight losses more heavily than gains of the same absolute size (if $\lambda = 1$, then Eq. (1) reduces to the normal standard gamble valuation function).

Therefore, Eq. (1) implies that the value of $x_2$ equates to the value of the perceived gains balanced against the perceived losses, relative the adopted reference point in the risky treatment arm of the gamble. It is worth noting that Bleichrodt et al. (2001) also considered the possible impact of probability transformation on standard gamble values. However, as noted above, Bleichrodt (1996; 2001) and also Oliver (2003) have observed that controlling for probability transformation appears to have a detrimental effect on the internal consistency of the standard gamble, and therefore probability transformation will not be considered in this article.

In agreement with Bleichrodt (2001), Spencer (2001) has argued that when coupled with diminishing marginal sensitivity (i.e. the observation that the marginal perceived impact of a change in outcome diminishes with the size of the outcome (Tversky & Kahneman, 1992)), loss aversion can explain the observed internal inconsistency in the standard gamble. Diminishing marginal sensitivity actually occurs with respect to both gains and losses. However, loss aversion could cause the individual to place particular emphasis on the failure outcome in the standard gamble. Therefore, the failure outcome may almost always be perceived as a *substantial* loss and, due to the principle of diminishing marginal sensitivity, the change in disutility when one substantial loss is replaced with another (i.e. when the failure outcome is altered during the chaining process) will be minimal. Therefore, loss aversion may ensure that the diminishing marginal sensitivity is most likely to transmit itself to the individual's responses when chaining to the failure outcome (Chilton & Spencer, 2001; Spencer, 2001). In sum, loss aversion could induce the individual to perceive the failure outcomes in the basic reference and chained to failure gambles to be too similar to ensure internal consistency (Bleichrodt, 2001; Spencer, 2001).

A further assumption of EU (and the standard gamble) that has often failed to hold in experimental settings is 'invariance' (e.g., Kahneman & Tversky, 1979; 1984; Tversky & Kahneman, 1981; 1986; 1992). Invariance requires that variations in the presentation of alternatives should not influence preference behaviour, *ceteris paribus* (Tversky & Kahneman, 1986). Violations of this assumption are commonly known as framing effects, and a number of studies have demonstrated that people's preferences over treatment options are often dependent upon whether the outcomes of treatment are framed (i.e. described) in terms of success or failure (McNeil, Pauker, Sox, & Tversky, 1982; O'Connor, Boyd, Tritchler, Kriukov, Sutherland, & Till, 1985; O'Connor, 1989). It is therefore quite plausible that values will be influenced by the way in which the standard gamble is framed, which, as with chaining, would be problematic as we would not know which frame generates values that are most commensurate with people's underlying strengths of preference for health states.

One possible way in which to decide which frame is most appropriate is to choose that which limits the level of internal inconsistency in the standard gamble. Tversky and Kahneman

(1991; 1992) have estimated $\lambda$ (re. Eq. (1)) to be approximately equal to two in both riskless and risky settings, although there is some evidence to suggest that the size of $\lambda$ is prone to manipulation. For example, Tversky and Kahneman (1991) state that loss aversion could vary across different settings, and Bleichrodt and Pinto (2002) report some empirical evidence that shows that loss aversion varies across different questions, let alone different frames. If loss aversion is malleable, and if loss aversion is indeed an explanation for the internal inconsistency in the standard gamble, then it may be possible to reduce the inconsistency through a relatively simple reframing of the standard gamble (i.e. by placing emphasis on treatment success and thus drawing attention away from $\lambda$).

### *Objectives*

The aims of the study reported in this article are twofold. The first objective is a relatively minor consideration and is simply to check if the answers that people give to standard gamble questions are influenced according to whether the gamble is framed in terms of success or failure. The expectation is that when emphasis is placed on treatment failure, people will become more focussed upon avoiding this outcome. Hence, the treatment arm of the standard gamble will appear less attractive, and higher values will be elicited than when emphasis is placed on treatment success. The second objective is the main focus of the article and is to test the internal consistency of the standard gamble in both the success and failure frames. As detailed above, it is plausible that loss aversion causes at least some of the internal inconsistency in the standard gamble. Placing emphasis on treatment success or failure is consistent with an attempt at varying the influence of loss aversion. When emphasis is placed on treatment success, we may expect the influence of loss aversion (i.e. $\lambda$ in Eq. (1)), and hence the observed internal inconsistency, to be less than when emphasis is placed on treatment failure. Thus, this study will address the issue of whether a relatively simple reframing of the standard gamble can improve its level of internal consistency.

## Methods

### *Respondents*

Thirty respondents, recruited via the London School of Economics e-mail network, participated in the study. The study comprised two interviews per respondent, who were paid £12 per interview and thus £24 in total. All of the respondents were either undergraduates (77%) or graduates (23%), 63% were female, 83% were aged 18-30 (the remainder were aged 31-45), 77% had a social science background (with a further 7% ‘humanities’, 7% ‘science’ and 10% ‘other’), and 77% were unfamiliar with decision theory.

### *Design*

As noted above, the study required the respondents to attend two interviews, spaced an average of two to three weeks apart. Each interview lasted between 15 and 30 minutes. The interviews were undertaken face to face (thus, the author administered 60 interviews in total), and in each interview the respondents were presented with a total of 17 questions. The interviews were divided into two parts. The study reported in this article is based upon only the first part, which, for each interview, comprised one practice question and ten main questions.

The respondents were required to answer the practice question at the beginning of each interview. The author took the respondents through a pre-answered question and the practice question. The respondents were free to ask questions during the practice session. After completing the practice question, the order in which the ten remaining questions were presented was randomised across respondents so as to reduce the possibility of ordering effects, and the respondents were required to answer these without asking any questions. The respondents were free to return to previous questions in order to revise their answers.

In one of the interviews the questions were framed in terms of success, and all of the questions were presented as health care contexts. The design of the questions mirrored the standard gamble format. A typical question posed in this interview is presented in Figure 2.

[Insert Figure 2]

The respondents were asked to imagine that they themselves face the choices presented in the questions. The three outcomes in each question, $x_1$, $x_2$ and $x_3$ (where, in Figure 2, $x_2$ = 10 years, $x_1$ = 40 years and $x_3$ = 0 years), were defined by length of life. This contrasts with the previous tests of the internal consistency of the standard gamble (Llewellyn-Thomas et al., 1982; Rutten-van Molken et al., 1995; Bleichrodt, 1996; 2001; Morrison, 1996; Spencer, 2001), where health outcomes were defined by the severity of the health state. Since some respondents may have difficulty in comprehending health states of varying severity, longevity is used so as to simplify the elicitation process and thus identify the extent to which anomalies in the standard gamble occur when it is stripped to its basic elements. The study simply requires the use of an outcome measure that is associated with positive marginal utility; it was assumed that all of the respondents would prefer more life in full health to less life in full health. Therefore, $x_1$, $x_2$ and $x_3$ could be set unambiguously at $x_1 \succ x_2 \succ x_3$, where $\succ$ means ‘is better than’. The maximum length of life on offer throughout the questionnaire was 40 years in good health. All of the respondents could reasonably expect to live for 40 more years.

In the interview from which the question in Figure 2 is taken, the successful outcome of treatment is emphasised. For example, the successful outcome is placed above the unsuccessful outcome, and the sentence, ‘So, if the treatment is successful, you will live for 40 years’, focuses on the treatment succeeding. Moreover, the respondents were asked for the minimum chance of *success* they would require in order for them to accept the treatment.

In the other interview, the question format was in all ways identical to that presented in Figure 2, with the important exception that the questions were framed in terms of the treatment failing. Thus, rather than indicating on the scale the minimum chance of success required, the respondents were asked to indicate the maximum chance of *failure* they would allow for them to accept the treatment. The question in the ‘failure frame’ that corresponds to that presented in Figure 2 is given in Figure 3.

[Insert Figure 3]

In the standard gamble, it is usual practice to attempt to elicit respondents’ indifference probabilities for p. However, when piloting the questions, it was found that people sometimes have difficulty in comprehending the concept of indifference in this study design. Therefore, the respondents’ respective stated minimum chance of success and maximum chance of failure are used as approximations of the probability that would render them indifferent between accepting the certainty, $x_2$, or the treatment.

In order to further reduce the possibility of ordering effects, 15 respondents (50%) answered the interview where the questions were framed in terms of treatment success first, and the other 15 respondents answered the interview where the questions were framed in terms of treatment failure first. Copies of the full questionnaires used in both interviews are available from the author on request.

*Tests*

The outcomes used in the questions are summarised in Table 1. As explained above, each question in the 'success frame' interview contained the same outcomes as the corresponding question in the 'failure frame' interview.

[Insert Table 1]

The first set of tests aims to ascertain if the respondents' answers significantly differed depending on whether the gamble is framed in terms of treatment success or failure (the 'framing tests'). Under the assumptions of EU, we would not expect a significant, systematic difference between the given probabilities for a question in the success frame interview and one minus the given probabilities for the corresponding question in the failure frame interview. Note that the stated probabilities, rather than the elicited values, are used in the framing tests. This is because in questions (qs.) 4 to 10, any difference in the *chained* values between frames may be entirely dependent on the differences in the values derived from qs. 1 to 3.

The second, more important, set of tests relates to the internal consistency of the standard gamble. Consider Table 1 in relation to the success frame interview. Assign a value of one to living for 40 years in full health (the best outcome in the study), and a value of zero to immediate death. The stated probabilities in qs. 1, 2 and 3 give the respondent's basic reference values for living for 10, 20 and 30 years in full health, respectively. Chained values can be elicited via qs. 4 and 5 for 10 years, qs. 6, 7 and 8 for 20 years, and qs. 9 and 10 for 30 years. The series of null hypotheses of perfect internal consistency between the basic reference and chained gambles are presented with the results in Table 3. Similar tests of internal consistency are undertaken with the answers given in the failure frame interview.

All tests of significance are calculated with Wilcoxon's matched-pairs signed-ranks test, a distribution-free test of the null hypothesis that two matched samples are drawn from identical populations. The tests used in this article are two-tailed at the 5% and the 1% levels of significance.

**Results**

The results of the framing tests are summarised in Table 2, and those of the tests of internal consistency are summarised in Table 3.

[Insert Tables 2 and 3]

The respondents’ answers significantly differed depending on whether the gambles were framed in terms of success or failure (Table 2). As expected, the stated probabilities in the success frame were generally of lower magnitude than one minus the stated probabilities in the failure frame (implying that higher values for any particular outcome would be elicited from the failure frame).

In the tests of internal consistency in the success frame (Table 3), the basic reference values did not significantly differ from the chained values when chaining to the success outcome; i.e. q.1 versus q.4, q.1 versus q.5 and q.2 versus q.6. When chaining to the failure outcome, the values elicited from the basic reference gambles did significantly differ from the chained values in three of the four comparisons. Where differences were observed, the tendency was for the chained values to exceed the basic reference values. Identical conclusions can be stated with respect to the tests of internal consistency in the failure frame (with the small exception that the difference between the values elicited from qs. 2 and 7 was strengthened from the 5% to the 1% level of significance).

As noted earlier, the respondents stated minimum required chances of treatment success and maximum allowed chances of treatment failure were used to approximate their indifference probabilities. Thus, for the former, the stated answers are likely to be at, or marginally above, the upper boundary of the range of indifference, and, for the latter, the stated answers are likely to be at, or marginally below, the lower boundary of the range of indifference. To serve as a sensitivity check, all tests were recalculated after multiplying all of the respondents’ minimum required chances of treatment success and maximum allowed chances of treatment failure by arbitrary factors of 0.98 and 1.02, respectively. By following this procedure, significant differences at the 1% level were observed in all ten framing tests (again with the failure frame implying higher values in all tests). The only observed alterations in terms of the results of the tests of internal inconsistency were that, in the success frame, there was no longer an observed significant difference between the values elicited from qs. 2 and 7 (the difference was very marginally above the 5% level of significance), but a significant difference was now observed between the values elicited from qs. 3 and 10 (at the 1% level). In the failure frame, a significant difference was also now observed between the values elicited from qs. 3 and 10 (at the 5% level). All significant inconsistencies remained in the direction of chained values exceeding basic reference values. Therefore, the overall impression is that the sensitivity analysis, if anything, very slightly strengthened the results.

## Discussion

The results of this study indicate a marked framing effect in the direction that was expected; i.e. answers were given that indicated that, compared to the success frame, significantly higher values would be elicited in the failure frame. A possible explanation for these results is that losses loomed larger when the failure outcome was emphasised.

In accordance with previous empirical findings, internal inconsistency was generally observed when the gamble was chained to the failure outcome. However, the degree of internal inconsistency was almost identical in the success and failure frames. Nonetheless, we cannot dismiss loss aversion as an explanation for the observed internal inconsistency. This is because the internal inconsistency in the success frame was itself quite pronounced and widespread when chaining to the failure outcome; three of the four tests that chained to the failure outcome generated significant levels of internal inconsistency. Given that only one of the four tests that chained to the failure outcome did not show internal inconsistency in the success frame, there was little opportunity for the failure frame to produce more marked levels of internal inconsistency when chaining to the failure outcome (even *if* loss aversion was greater in the failure frame). It is therefore possible that substantial loss aversion was induced in both the success and failure frames, and that loss aversion could not be *sufficiently* manipulated to improve internal consistency through a relatively simple reframing of the standard gamble.

In both frames, the question that involved evaluating 30 years in good health (the least 'severe' health state evaluated in the study) against a treatment offering a chance of 40 years or 20 years in full health, did not generate statistically significant internal inconsistency (see Table 3). Other researchers have also found less pronounced internal inconsistency in the evaluation of minor, as opposed to severe, health states (Morrison, 1996; Spencer, 2001). Relatively minor health states may generate less internal inconsistency because their basic reference value will be close to one (i.e. as noted in the introduction, the standard gamble may be an insufficiently sensitive instrument for the valuation of minor health states), leaving relatively little scope for higher chained values (Bleichrodt, 1996; Spencer, 2001).

There are a number of noteworthy caveats in the study. In hindsight, one of the problems with the methods is that they do not include a full range of health outcomes across the utility continuum. In the basic reference gambles, the respondents generally gave a high valuation of 10, 20 and 30 years in full health. This could have been accentuated by the choice of length of life as the health outcome, in that the respondents may have had a high discount rate over life years. However, a high discount rate over life years cannot explain the observed internal inconsistency. For example, consider the following basic reference and chained gambles:

q.2: $v(20 \text{ years}) = p_2 \cdot v(40 \text{ years}) + (1- p_2) \cdot v(0 \text{ years})$ [Basic reference gamble]

q.8: $v(20 \text{ years}) = p_8 \cdot v(40 \text{ years}) + (1- p_8) \cdot v(10 \text{ years})$ [Chained gamble]

As detailed earlier in this article and elsewhere (Spencer, 2001), people may not perceive the failure outcome in the chained gamble to be sufficiently better than the failure outcome in the basic reference gamble. This will lead to $p_2$ being insufficiently larger than $p_8$ to ensure internal consistency, an occurrence that has thus far been attributed to loss aversion. However, if an individual has a very high discount rate over life years, then the value that they place on 10 years in full health may be similar to that which they place on immediate death. Such an occurrence would also lead to an insufficiently large $p_2$ relative to $p_8$, as the individual would perceive q.2 and q.8 as almost identical. However, 24 of the 30 respondents in the success frame, and 28 in the failure frame, gave a basic reference value of $\geq 0.7$ for 10 years in full health, and therefore it is apparent that most people *do not* intrinsically value 10 years in full health and immediate death at similar levels.

Two further caveats of the study ought to be noted. The first is that the health outcomes employed were unrealistically simple. It is perhaps difficult for most individuals to imagine that they can take a treatment for a chance of living for 10, 20, 30 or 40 years in full health and then, abruptly, die, and we never face situations where we know that we will live for a specific number of years for certain (Johannesson, Jönsson, & Karlsson, 1996). However, as mentioned in the methods section, simple health outcomes defined by life expectancy were used so as to reduce the standard gamble to its basic elements. Second, the results reported in this article were derived from a relatively small sample. Therefore, care ought to be taken to avoid over-interpretation of the results, and further studies with larger sample sizes using more realistic health outcomes in real clinical settings would be interesting and useful.

## Conclusion

To conclude, framing in terms of treatment success or failure clearly influences values, which in itself is problematic in that we are left unsure as to which values most accurately reflect strengths of preference for health states. Moreover, a relatively simple reframing of the standard gamble to attempt to draw attention away from the failure outcome and hence reduce the influence of loss aversion failed to improve the level of internal consistency. An important, seemingly robust, concern is that the standard gamble suffers from significant levels of internal inconsistency when chaining to the failure outcome, which demands further methodological development of this elicitation instrument.

Of course, the non-effect of reframing the standard gamble on its level of internal consistency may have been observed because loss aversion is not, after all, a cause of the internal inconsistency. However, an alternative and perhaps more likely explanation is that the fundamental construct of the standard gamble induces a focus on the failure outcome and hence a significant degree of loss aversion, irrespective of how the instrument is framed. It thus remains possible that loss aversion is one (though probably not the sole) cause of the observed internal inconsistency. If we accept that this explanation is plausible, further tests of internal consistency ought to be undertaken (1) after mitigating loss aversion with a more sophisticated internalisation of this psychological process in the standard gamble valuation procedure (Bleichrodt, Pinto, & Wakker, 2001; Oliver, 2003), and/or (2) on alternatives to the standard gamble that intuitively induce less loss aversion (McCord & de Neufville, 1986). These tasks are beyond the scope of this article, but both offer scope for interesting and potentially important research.

## Appendix

Using a basic reference standard gamble format, assume that an individual is asked to value two health states, 'paralysed (from the waist down)' and 'deaf'. For this hypothetical individual, a value of 0.4 is elicited for the health state, 'paralysed' and a value of 0.6 is elicited for the health state, 'deaf'. It follows that if the individual concords perfectly with the axioms of EU, then she ought to state a probability of 0.3333 if deaf is evaluated against full health and paralysed; i.e. $0.6 = p + (1-p)0.4$, and therefore $p = 0.3333$.

Assume that the individual does not concord with EU. Assume instead that she evaluated the standard gambles in the manner expressed in Eq. (1). That is:

$$v(x_2) = v(x_2) + p(v(x_1)-v(x_2)) - \lambda(1-p)(v(x_2)-v(x_3))$$

$$\Rightarrow v(x_2) = (pv(x_1) + \lambda(1-p)v(x_3))/(p + \lambda(1-p)) \qquad \text{(A1)}$$

Given that the individual gives respective probabilities of 0.4 and 0.6 when valuing paralysed and deaf in the basic reference EU standard gamble, the second and third columns in Table A list the values elicited with Eq. (A1) for paralysed and deaf at varying levels of loss aversion, with the (arbitrary) levels of loss aversion ($\lambda$) listed in the first column of the table. It is worth noting that the greater the $\lambda$, the more biased are the EU standard gamble values. For example, if the individual has a $\lambda = 1.25$, then when stating a probability = 0.4 for the health state 'paralysed' in a basic reference gamble, her underlying loss-adjusted value of this health state = 0.3478; if she has a $\lambda = 2.5$, her loss-adjusted value of this health state is only 0.2105.

[Insert Table A]

If the individual displays loss aversion when answering the standard gamble in the manner hypothesised in this article, then when 'deaf' is evaluated against full health and 'paralysed', internal consistency will require her to express a probability, $p^*$, given by:

$$p^* = \lambda(v(x_3) - v(x_2))/(v(x_2) - \lambda v(x_2) - v(x_1) + \lambda v(x_3)) \qquad \text{(A2)}$$

where Eq. (A2) is derived from Eq. (A1).

The $p^*$ at the different specified levels of $\lambda$ listed in Table A are given in the fourth column of the table. However, if it is incorrectly assumed that the individual concords with EU, then the indirect chained value of 'deaf' will be calculated with the normal EU valuation formula. Moreover, in these circumstances, it will also be incorrectly assumed that the individual's stated probability in the basic reference gamble used to evaluate the health state 'paralysed' (i.e. 0.4, in this example) accurately represents her value of this health state, and thus this value will be substituted into the calculation for the chained value of 'deaf'. These biased calculations for each specified level of loss aversion are listed in the fifth column of the table.

Therefore, for example, it can be observed in Table A that if the individual weighs losses precisely twice as much as gains of the same absolute size (i.e. $\lambda = 2$), then when stating a probability of 0.4 and 0.6 in the basic reference gambles when evaluating 'paralysed' and 'deaf' respectively, their underlying loss-adjusted values of these health states should be calculated at 0.25 and 0.4286. Consequently, to ensure internal consistency, the individual should state a probability of 0.3847 when ‘deaf’ is valued against full health and ‘paralysed’. If it is incorrectly assumed that this individual concords perfectly with EU, her basic reference value for deaf will be calculated at 0.6, and her chained value will be calculated at 0.6308, violating internal consistency in the direction that has been observed in the literature. As can be observed in Table 5, higher levels of loss aversion will render it more likely that any particular individual will demonstrate internal inconsistency when chaining to the failure outcome in the gamble, if the EU valuation procedure is followed.

The internal consistency-related hypothesis in this article is that more people will demonstrate at least some loss aversion, and that those that do demonstrate loss aversion will demonstrate it at higher levels, when the standard gamble is framed in terms of failure rather than success. This is the rationale for the a priori expectation that internal inconsistency will be more significant and systematic in the failure frame, and indeed that by drawing the focus away from the failure outcome, the observed internal inconsistency may be removed to any significant and systematic extent if one were to place a strong emphasis on the success outcome.

**Acknowledgements**

This study was part-funded by the Office of Health Economics and the article has benefited greatly from comments received from Mike Jones-Lee, Anne Spencer, Steve Birch and two anonymous referees. The author is also grateful for ESRC award number R00429834596. The content of the article, and any mistakes therein, are the sole responsibility of the author.

## References

Bleichrodt, H. (1996). *Applications of utility theory in the economic evaluation of health care*. PhD Thesis, Rotterdam: Erasmus University.

Bleichrodt, H. (2001). Probability weighting in choice under risk: an empirical test. *Journal of Risk and Uncertainty*, 23, 185-198.

Bleichrodt, H. & Pinto, J. L. (2002). Loss aversion and scale compatibility in two-attribute trade-offs. *Journal of Mathematical Psychology*, 46, 315-337.

Bleichrodt, H., van Rijn, J., & Johannesson, M. (1999). Probability weighting and utility curvature in QALY based decision making. *Journal of Mathematical Psychology*, 43, 238-260.

Bleichrodt, H., Pinto, J. L., & Wakker, P. P. (2001). Using descriptive findings of prospect theory to improve prescriptive applications of expected utility. *Management Science*, 47, 1498-1514.

Camerer, C., & Ho, T. (1994). Violations of the betweenness axiom and non-linearity in probability. *Journal of Risk and Uncertainty*, 8, 167-196.

Chilton, S., & Spencer, A. (2001). Empirical evidence of inconsistency in standard gamble choices under direct and indirect elicitation methods. *Swiss Journal of Economics and Statistics*, 137, 65-86.

Enemark, U. (1994). *Alternatives to expected utility theory in clinical decision making: a review of prospect theory, regret theory and disappointment theory.* Studies in Health Economics 5, Lund: Lund University.

Eraker, S. E., & Sox, H. C. (1981). Assessment of patients' preferences for therapeutic outcomes. *Medical Decision Making*, 1, 29-39.

Fischhoff, B. (1983). Predicting frames. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 9, 103-116.

Fishburn, P. C., & Kochenberger, G. A. (1979). Two-piece von Neumann-Morgenstern utility functions. *Decision Sciences*, 10, 503-518.

Hershey, J. C., & Schoemaker, P. J. H. (1980). Risk taking and problem context in the domain of losses: an expected utility analysis. *Journal of Risk and Insurance*, 47, 111-132.

Howell, D. C. (1989). *Fundamental statistics for the behavioral sciences.* Belmont, California: Duxbury Press.

Johannesson, M., Jönsson, B., & Karlsson, G. (1996). Outcome measurement in economic evaluation. *Health Economics*, 5, 279-296.

Jones-Lee, M. W., Loomes, G., O'Reilly, D., & Philips, P. (1993). *The value of preventing non-fatal road injuries: findings of a willingness-to-pay national sample survey*. Working Paper WP/SRC/2. London: Transport Research Laboratory.

Jones-Lee, M. W., Loomes, G., & Philips, P.R. (1995). Valuing the prevention of non-fatal road injuries: contingent valuation vs. standard gambles. *Oxford Economic Papers*, 47, 676-695.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, 47, 263-291.

Kahneman, D., & Tversky, A. (1984). Choices, values and frames. *American Psychologist*, 39, 341-350.

Llewellyn-Thomas, H., Sutherland, H. J., Tibshirani, R., Ciampi, A., Till, J. E., & Boyd, N. F. (1982). The measurement of patients' values in medicine. *Medical Decision Making*, 2, 449-462.

McCord M., & de Neufville, R. (1986). "Lottery equivalents": reduction of the certainty effect problem in utility assessment. *Management Science*, 32, 56-60.

McNeil, B. J., Pauker, S. G., Sox, H. C., & Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England Journal of Medicine*, 306, 1259-1262.

Morrison, G. C. (1996). *An assessment and comparison of selected methods for eliciting quality of life valuations from the public.* PhD Thesis, York: University of York.

O'Connor, A. M. (1989). Effects of framing and level of probability on patients' preferences for cancer chemotherapy. *Journal of Clinical Epidemiology*, 42, 119-126.

O'Connor, A. M. C., Boyd, N. F., Tritchler, D. L., Kriukov, Y., Sutherland, H., & Till, J. E. (1985). Eliciting preferences for alternative cancer drug treatments. *Medical Decision Making*, 5, 453-463.

Oliver, A. (2003). The internal consistency of the standard gamble: tests after adjusting for prospect theory. *Journal of Health Economics*. Forthcoming.

Payne, J. W., Laughhunn, D. J., & Crum, R. (1980). Translation of gambles and aspiration level effects in risky choice behavior. *Management Science*, 26, 1039-1060.

Rutten-van Molken, M. P. M. H., Bakker, C. H., van Doorslaer, E. K. A., & van der Linden, S. (1995). Methodological issues of patient utility measurement. *Medical Care*, 33, 922-937.

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

Torrance, G. W. (1986). Measurement of health state utilities for economic appraisal. *Journal of Health Economics*, 5, 1-30.

Torrance, G. W., & Feeny, D. (1989). Utilities and quality-adjusted life years. *International Journal of Technology Assessment in Health Care*, 5, 559-575.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211**,** 453-458.

Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business*, 59, S251-S278.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: a reference dependent model. *Quarterly Journal of Economics*, 107, 1039-1061.

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

Wu, G., & Gonzalez, R. (1996). Curvature of the probability weighting function. *Management Science*, 42, 1676-1690.

Table 1: Summary of outcomes used in the questionnaire

| | Years in full health | | |
|---|---|---|---|
| | $x_2$ | $x_1$ | $x_3$ |
| Practice 1 | 10 | 40 | 0 |
| | | | |
| Question 1 | 10 | 40 | 0 |
| Question 2 | 20 | 40 | 0 |
| Question 3 | 30 | 40 | 0 |
| | | | |
| Question 4 | 10 | 20 | 0 |
| Question 5 | 10 | 30 | 0 |
| | | | |
| Question 6 | 20 | 30 | 0 |
| Question 7 | 20 | 30 | 10 |
| Question 8 | 20 | 40 | 10 |
| | | | |
| Question 9 | 30 | 40 | 10 |
| Question 10 | 30 | 40 | 20 |

Table 2: Summary of results: tests of the framing effect

| | | | |
|---|---|---|---|
| q.1: | ** | q.6: | * |
| q.2: | * | q.7: | ** |
| q.3: | ** | q.8: | ** |
| q.4: | ** | q.9: | * |
| q.5: | ** | q.10: | ** |

Note: * indicates a significant difference at 5%; ** indicates a significant difference at 1%

Table 3: Summary of results: tests of internal consistency

| | Hypothesis | Success frame | Failure frame |
|---|---|---|---|
| 10 years | | | |
| q.1 v q.4: | $p_1 = p_4p_2$ | ns | ns |
| q.1 v q.5: | $p_1 = p_5p_3$ | ns | ns |
| | | | |
| 20 years | | | |
| q.2 v q.6: | $p_2 = p_6p_3$ | ns | ns |
| q.2 v q.7: | $p_2 = p_7p_3 + (1-p_7)p_1$ | * | ** |
| q.2 v q.8: | $p_2 = p_8 + (1-p_8)p_1$ | ** | ** |
| | | | |
| 30 years | | | |
| q.3 v q.9: | $p_3 = p_9 + (1-p_9)p_1$ | ** | ** |
| q.3 v q.10: | $p_3 = p_{10} + (1-p_{10})p_2$ | ns | ns |

Note: In the success frame, $p_i$ given in the hypotheses is the respondent's stated probability in q.i (in the failure frame, pi equates to one minus the respondent's stated probability in q.i); ns indicates that there was not a significant difference between the elicited values; * indicates a significant difference at 5%; ** indicates a significant difference at 1%.

Table A: Example of the influence of loss aversion on internal inconsistency

| $\lambda$ | v(paralysed) | v(deaf) | $p^*$ | $p^* + (1-p^*)0.4$ |
|---|---|---|---|---|
| 1 (=EU) | 0.4 | 0.6 | 0.3333 | 0.6 |
| 1.25 | 0.3478 | 0.5455 | 0.3522 | 0.6113 |
| 1.50 | 0.3077 | 0.5 | 0.3659 | 0.6195 |
| 1.75 | 0.2759 | 0.4615 | 0.3762 | 0.6257 |
| 2.00 | 0.25 | 0.4286 | 0.3847 | 0.6308 |
| 2.25 | 0.2286 | 0.4 | 0.3913 | 0.6348 |
| 2.50 | 0.2105 | 0.375 | 0.3969 | 0.6381 |

Figure 1: The internal consistency of the standard gamble

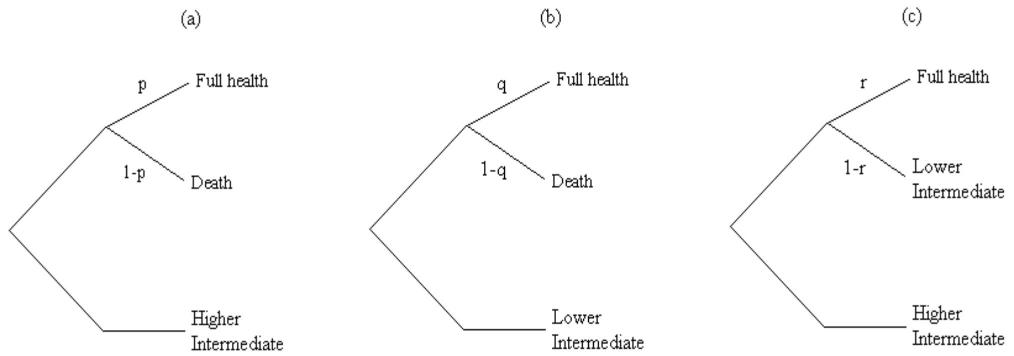


Figure 2: A typical question in the success frame

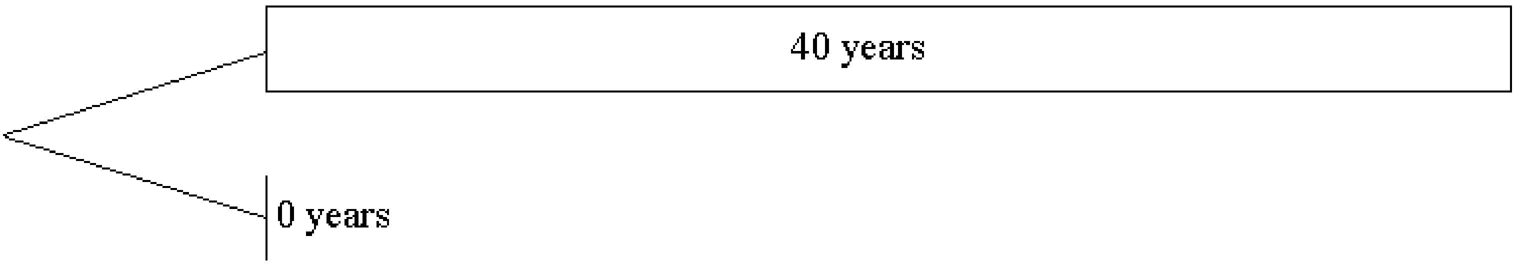
Imagine that you go to your doctor for a routine medical examination. To your surprise, your doctor informs you that you have an unusual health condition. In this condition your doctor informs you that, without treatment, you will live for 10 more years in good health, and then you will die.

However, your doctor also informs you that there is a treatment for your condition, which, if taken, would give you a chance of living for 40 years in good health before death. However, there is also a chance that the treatment would kill you immediately.

So, your two options are:

1. Do not take the treatment and live for 10 years.


2. Take the treatment for a chance of living for 40 years and risk the chance of immediate death.




So, if the treatment is successful, you will live for 40 years.

Your doctor tells you that the size of the chance that the treatment will succeed is not known.

Please indicate on the scale below **the minimum chance of success** you would require for you to accept the treatment.

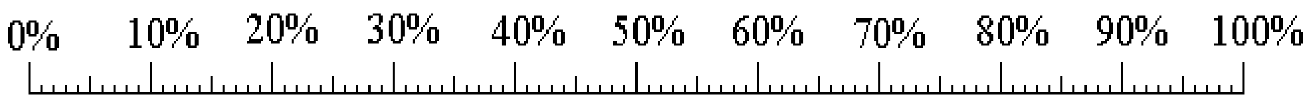


Figure 3: A typical question in the failure frame

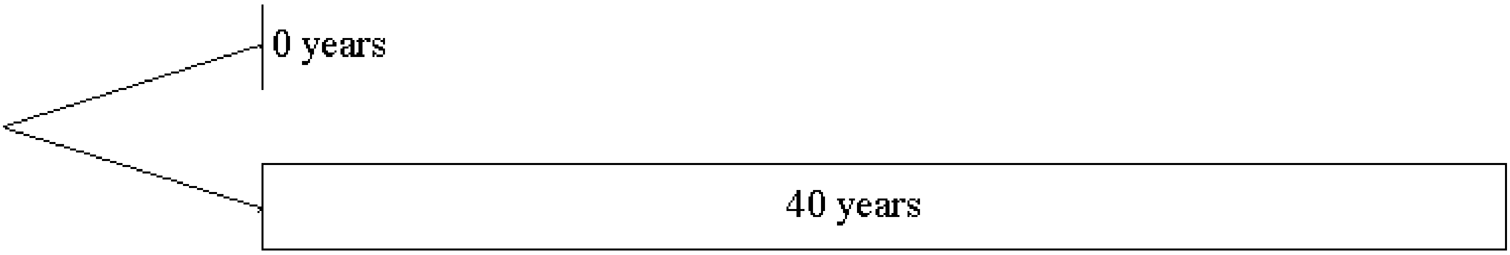
Imagine that you go to your doctor for a routine medical examination. To your surprise, your doctor informs you that you have an unusual health condition. In this condition your doctor informs you that, without treatment, you will live for 10 more years in good health, and then you will die.

However, your doctor also informs you that there is a treatment for your condition. Though there is a chance that the treatment would kill you immediately, there is also a chance that it would give you 40 years in good health before death.

So, your two options are:

1. Do not take the treatment and live for 10 years.


2. Take the treatment and risk the chance of immediate death for a chance of living for 40 years.




So, if the treatment fails, you will die immediately.

Your doctor tells you that the size of the chance that the treatment will fail is not known.

Please indicate on the scale below **the maximum chance of failure** you would allow for you to accept the treatment.