Biased risk perception in health state utility measurement: a new type of framing effect

Abellán Perpiñán, JM (Email: dionisos@um.es)

Martínez Pérez, JE

Universidad de Murcia, Departamento de Economía Aplicada

Campus de Espinardo. 30100. Murcia.

Pinto Prades, JL

Universidad Pompeu Fabra, Departamento de Economía y CRES

Abstract

This paper tests whether logically equivalent risk formats can lead to different health state utilities elicited by means of the standard gamble (SG) method. We compare SG utilities elicited when probabilities are framed in terms of frequencies with respect to 100 people in the population (i.e., X out of 100) with SG utilities elicited for frequencies with respect to 1,000 people in the population (i.e., Y out of 1,000). We find that utilities were significant higher when success and failure probabilities were framed as frequencies type “Y out of 1,000” rather than as frequencies type “X out of 100”. This violation of invariance holds with independence on the severity of the health status and the statistical procedure used to test significance (bootstrap, Wilcoxon Mann-Whitney test). This framing effect can be explained by the joint effect of a heuristic we name the *numerator effect* and the phenomenon of loss aversion, formalized by Prospect Theory.

KEYWORDS: framing effect, risk format, standard gamble, health state

Introduction

The term framing is currently used to describe the wording or presentation of logically equivalent information. A framing effect (Tversky and Kahneman, 1981) arises when alternative framing of formally identical problems lead to different choices. The available evidence shows that framing is a reliable phenomenon to affect both hypothetical and real treatment choices of patients and physicians (see Edwards et al., 2001, for a review).

Empirical evidence encompasses several types of framing effects. First, preference reversals can occur when equivalent risk information is presented in a negative or a positive frame, or in a gain vs loss frame (Eraker and Sox, 1981; O'Connor et al., 1985; O'Connor, 1989; Banks et al., 1995; Redelmeier et al., 1993; Llewellyn-Thomas et al., 1995; Gurm and Litaker, 2000; Amstrong et al., 2002). A classic example of this type of framing effect is the preference reversal between surgery versus radiation described by McNeil et al. (1982). Patients to whom surgical mortality was presented as a 10% probability of dying were less willing to choose surgery than patients presented with surgical mortality as a 90% probability of surviving. Second, quite different choices over alternative treatments can emerge when information is presented as relative risk, absolute risk, or the number-needed-to treat (Forrow et al., 1992; Malenka et al., 1993; Bucher and Weinbacher, 1994; Hux et al., 1994; Sarfati et al., 1998; MtGettigan et al., 1999). For example, around 90% of participants in a study conducted by Hux and Naylor (1995) preferred to have a medicine when effectiveness was expressed as a reduction of relative risk, whereas the percentage was only 30% when effectiveness was described in terms of the number-needed-to treat format. Third, some studies have found that numerically equivalent risk formats (e.g. frequency versus percentage) can lead to inconsistent preferences. Participants in an experiment conducted by Yamagishi (1997a) rated a disease that kills 1,286 people out of every 10,000 as more dangerous than one that kills 24.14 out of every 100. Slovic et al. (2000) found that clinicians were more likely to keep a mental patient in the hospital if the risk that the patient would commit an act of violence after being discharged had been communicated in a frequency format (e.g. 20 out of 100 patients) as compared to a percentage format (e.g. 20% likely). Schapira et al. (2004) reported disagreement in breast cancer risk perceptions when measured by a frequency and a percentage scale.

This paper is concerned with the latter type of framing effect we have just mentioned. Specifically, we test whether logically equivalent risk formats can lead to different health state utilities elicited by means of the standard gamble (SG) method. Although there exist some evidence of the influence of the gain vs loss framing on risky elicitation methods (Stalmeier and Bezembinder, 1999; Bernstein et al., 1999), to date no study has addressed the issue if utilities elicited by the SG are invariant to the format in which probabilities are presented. We think this question is important because the SG, one of the techniques currently used to elicit health state utilities, assumes that people can accurately understand risk information in such a way that preferences should be invariant to the risk format. If preferences are not invariant to the risk format then SG measurements could be arbitrary and medical decisions based on SG utilities could be biased.

The paper is structured as follows. In section 2 we describe the invariance tests to the risk format used in the study. Probabilities in our study were described in terms of relative frequencies (e.g., X out of 100 people). The use of frequency formats is supporting by those authors (Gigerenzer and Hoffrage, 1995; Hoffrage and Gigerenzer, 1998; Hoffrage et al., 2000) that have suggested that, for diagnostic inference, such as the determination of pre- and post-test probabilities, frequency formats are easier for people to use than percentage formats. In section 3 we hypothesize that SG utilities elicited when probabilities are framed in terms of frequencies with respect to 1,000 people in the population will be higher than SG utilities elicited for frequencies with respect to 100 people in the population. This prediction is based on the joint effect of a mental shortcut we coin as the numerator effect and the loss aversion bias described in Prospect Theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992). Section 4 describes an experiment designed to test invariance to the risk format with the SG. Questions were administered to a convenience sample (students). Four EQ-5D health states were described as chronic conditions. This section also shows results from the experiment. Discussion closes the paper.

2. Invariance tests

Let (Q, RL) be a chronic health state (i.e., a permanent condition until dead), denoting rest of life, RL, in health state Q. From now on FH stands for full health and D stands for death. Let [(FH, RL), X/100; D, X*/100] denote the gamble that gives outcome (FH, RL) with probability "X out of 100" and death with probability "X* out of 100". The standard gamble asks for probabilities X/100 and X*/100 that yield indifference between the gamble [(FH, RL), X/100; D, X*/100] and a sure outcome (Q, RL). If it assumed the descriptive validity of expected utility then the SG utility of Q is computed as X/100. The same holds for probabilities expressed as frequencies with respect to 1,000 people in the population.

To test invariance of the SG, we splitted a convenience sample into two groups. Then we asked participants in one group to set probabilities that yield indifference between the gamble [(FH, RL), X/100; D, X*/100] and a sure outcome (Q, RL). Overall, we asked four SG questions, one for each of the health states we selected. Hence, we performed four invariance tests. The same questions were also asked to participants in the other group, but probabilities were framed as frequencies Y/1,000 and Y*/1,000. The comparison for each health state between SG utilities elicited from one group (henceforth, U_{SG100}) and SG utilities elicited from the other group ($U_{SG1,000}$) provides four invariance tests of the SG method.

3. The numerator effect

It has been demonstrated that people respond differently to equivalent formats of relative frequency information that are presented in superficially different ways. For example, Denes-Raj and Epstein (1994) showed that, when offered a chance to win $1 by drawing a red jelly bean from an urn, people often chose to draw from a bowl containing a greater absolute number, but a smaller proportion, of read beans (9 out of 100) than from a bowl with fewer red beans but a better chance of winning (1 out of 10). When participants were asked to justify their choice, they admitted that this choice went contrary to what a rational individual should do, but they felt they had a better chance when there were more red beans.

Slovic et al. (2004) characterize this tendency as a manifestation of a mental strategy of "imaging the numerator" (the number of red beans) and "neglecting the denominator" (the number of beans in the bowl). According to Slovic et al., images of winning beans convey positive affect that motivates the choice of the bowl with the greater absolute number of read beans. This affective mechanism was coined by Slovic et al. (2002) as the *affect heuristic*.

A similar reasoning is argued to explain that Slovic et al. (2000) found 41% of clinicians refused to discharge a mental patient when patient's violence risk was communicated as "20 out of every 100 patients similar to Mr. Jones are estimated to commit an act of violence", but only 21% refused to discharge the patient when risk was given as "patients similar to Mr. Jones are estimated to have a 20% chance of committing an act of violence". In this case, as it is easier to visualize frequencies than percentages, "imaging the numerator" (i.e., 20 mental patients commit violent acts) can lead to clinicians are more fearful when risk is communicated in a frequency format, and to react to that fear keeping the patient in the hospital. Consistent with the affect heuristic, images of violent patients convey negative affect that motivates conservative risk management decisions.

Whereas Slovic et al. propose an affect heuristic to explain the higher impact of relative frequencies with a larger numerator, Yamagishi (1997a) suggest a combination of two cognitive mechanisms, namely, *anchoring and adjustment* (Tversky and Kahneman, 1974) and *base-rate neglect* (Kahneman and Tversky, 1973), respectively. As noted in introduction, Yamagishi found that a risk was judged as more serious when the deaths were expressed by a larger numerator (e.g., cancer killing 1,286 out of 10,000) than a smaller numerator (e.g., cancer killing 24.14 out of 100). Yamagishi argues that this result can be explained because people use the numerator as an anchor to do subsequent judgment and, simultaneously, tend to reject the base rate because it is cancer, not the denominator of the relative frequency, that kills patients. As 1,286 is higher than 24.14 then a risk of 1,286 out of 10,000 is perceived as more dangerous than a risk of 24.14 out of 100, even though the former probability (12.86%) is lower than the latter one (24.84%).

In sum, regardless perception of relative frequencies is motivated by affective or cognitive mechanisms, the fact is that people seem be simultaneously sensitive to the rote frequency (the numerator) and insensitive to the total

number of population (the denominator), being the final effect that a frequency “X out of 100” will be judged as a lower probability than a frequency “Y out of 1,000”, being $X < Y$, even though both risks are numerically equivalent (i.e., $x\% = y\%$). We refer to individual’s tendency of focusing on the numerator of a relative frequency and neglecting the denominator as the *numerator effect*.

To set our hypothesis we assume that risk perceptions when a frequency mode is used are guided by the heuristic of the numerator effect. In addition, we assume that SG measurements are affected by loss aversion. This bias predicts that losses are weighted more heavily than gains. Losses and gains are evaluated with respect to a reference point. Empirical evidence (Hersey and Schoemaker, 1985; Bleichrodt et al., 2001; Bleichrodt et al., 2005) seems to support that the outcome given in advance in a SG question (i.e., the sure outcome) is taken as a reference point, with respect to which the best outcome in the gamble is judged as a gain and the worst outcome in the gamble is judged as a loss.

In the invariance test of the SG method, the individual sets X/100 and X*/100 as indifference probabilities between the gamble [(FH, RL), X/100; D, X*/100] and the sure outcome (Q, RL). Hence, $U(Q)_{SG100} = X/100$. The absolute number of deaths with respect to 1,000 people in the population required to obtain the same SG utility would be $Y^* = 10 \cdot X^*$, given that $Y^*/1{,}000 = [10 \cdot X^*/10 \cdot 100]$. However, if it assumed that the individual perceives the gain from (Q, RL) to (FH, RL) with probability Y/1,000 as not sufficient to offset the loss from (Q, RL) to death with probability Y*/1,000 (i.e., if the individual is loss averse), and it is also assumed that the individual focuses on the numerator of relative frequencies, neglecting the denominator, then $10 \cdot X^*$ will be perceived as too many deaths in such a way that the individual will set $Y^* < 10 \cdot X^*$ in order to reach indifference. Thus, $U(Q)_{SG1,000} = Y/1{,}000 > X/100 = U(Q)_{SG100}$. This is the hypothesis we tested in our experiment.

4. The experiment

- **Sample**

The participants were 200 economics students at the University of Murcia (Spain). They were paid 6 Euro. Responses were collected in personal interview sessions. Prior to the actual experiment, we tested the questionnaire in several pilot sessions.

- **Health states**

We used the EQ-5D health states 22111, 11222, 22222, and 23222. These states are described in Table 1. Throughout the experiment, the health states were labeled health state X, W, Z, and Y, respectively.

[Insert table 1 about here]

Given the ordinal structure of the component dimensions in the EuroQol descriptive system, some states are logically ordered with respect to others. With the states used here, five such comparisons are possible. It would be expected that 22222 should be given a higher utility than 23222 because it is better on at least one dimension and no worse on any of the other dimensions. In the same way, it would be expected this ordinal consistency for comparisons between 22111 vs 22222, 22111 vs 23222, 11222 vs 22222, and 11222 vs 23222. For comparison between 22111 and 11222 there is no a priori expectation of this kind.

- **Design**

The experiment was run on a computer. Computers made easier to use visual aids and to apply choice-based procedures to elicit utilities. In addition, the pilot sessions showed that people found that the computer assisted personal interview was regarded as a user-friendly procedure.

To avoid anchoring biases and not to overburden the participants we splitted the total sample into two groups of 100 subjects each. One group answered four SG questions (one per health state) in which probabilities were framed in terms of frequencies with respect to 1,000 people in the population, while the other group answered the same questions for probabilities framed as frequencies with respect to 100 people in the population. To avoid order effects, the computer varied the order in which the different SG questions were asked. So as to minimize response errors, subjects had to confirm the elicited indifference value after each question. As a preliminary task, participants were asked to rate the health states on a visual analogue scale (VAS), with 100 (best imaginable

health state) and 0 (worst imaginable health state) as endpoints. The principal objective of the VAS was to familiarize participants with the health state descriptions.

Recruitment of participants took place one week before the actual experiment started. At recruitment, participants were handed a practice question. Participants were asked to answer this practice question at home. This procedure intended to familiarize participants with the SG questions and avoids problems of construction of preferences. Prior to the start of the experiment, subjects were asked to explain their answer to the practice question. When we were not convinced that a subject understood the task, we explained it again until we were convinced that he understood the task.

The formulation of SG questions was always the same, regardless the frequency format used. For example, in the case of health state X, the wording of the question was as follows:

> *Suppose that you are experienced health state X. If you do not receive treatment you will remain in that health state for the rest of your life. However, you can receive a medical treatment (ALPHA treatment) thanks to which, if it is successful, you will recovery full health. Nevertheless, ALPHA treatment can also fail. If it occurs you will dead.*
> *Next, according to the different probabilities of success and failure we are going to display, you should decide if you prefer ALPHA treatment or not.*

A choice-based elicitation procedure was applied to elicit utilities. Previous studies (Bostic et a., 1990; Luce, 2000) have found that choice-based procedures are less likely to give inconsistencies in preferences than matching procedures. Frequencies were displayed using human figures. It seems (Schapira et al., 2001) that frequency formats illustrated with human figures are attached to attributes of ease interpretation, simplicity, and ability to convey a meaningful message. Figures 1 and 2 illustrate the way indifferences were obtained.

[Insert figures 1 and 2 about here; visual aid]

Suppose that for the probabilities displayed in figure 1, the individual prefer treatment ALFA to no treatment. Next the computer would display a new choice (figure 2), where probabilities of success and failure would be 10 in 100 and 90 in 100 respectively. Suppose that for these probabilities the individual prefer the sure outcome rather than

treatment ALFA. Then the computer would vary probabilities of success and failure from 91/100 to 94/100 and from 6/100 to 9/100 respectively until the individual was indifferent between the two treatments.

- **Data analysis and statistical methods**

As normality was rejected by the Kolmogorov-Smirnov test with Lilliefors correction ($p < 0.0001$), we used nonparametric procedures in order to test significance of differences between SG utilities elicited for frequencies with respect to 100 people in the population and SG utilities elicited with respect to 1,000 people in the population.

First, we used bootstrapping to determine asymptotically distribution-free confidence intervals for differences between mean SG utilities. Specifically, we estimated bias-corrected-and-accelerated (BCA) confidence intervals. Properties of BCA confidence intervals are summarized, for example, by Efron and Tibshirani (1993). To estimate BCA confidence intervals we repeatedly (1,000 repetitions) analyze subsamples of the data. Each subsample is a random sample with replacement from the full sample. Next, means were computed and finally confidence interval for differences between means were determined. We rejected equality between SG means utilities when 0 is not included within bounds of the confidence interval. The same analysis was also done for trimmed means. We calculated trimmed means (Barnett and Lewis, 1995) by eliminating values whose difference with respect the mean exceeded the standard deviation in more than β times. Overall, we obtained bootstrap confidence intervals for $\beta = 0, 0.5, 1, 1.5, 2, 2.5$.

Second we also tested whether the two distributions of SG utilities differed with respect to median. To that end we used the Wilcoxon Mann-Whitney test. This is one of the most powerful of the non-parametric tests for comparing two independent samples (Hollander and Wolfe, 1999).

- Results

Table 2 shows means, medians and standard deviations (SD) corresponding to VAS and SG valuations for the full sample (N = 200). Both methods yield logically consistent rankings in the sense that state Y receives a higher

value than state Z, and both Y and Z receive lower values than states X and W. In addition, SG utilities are higher than VAS valuations for all health states.

[Insert table 2 about here]

Table 3 displays descriptive statistics referred to SG utilities for each one of the two groups (N = 100) we splitted the full sample. The picture that emerges from the table is clear. Utilities elicited by the SG method from the group in which frequencies were described as “Y out of 1,000” are consistently higher than utilities elicited from that group in which frequencies were described as “X out of 100”. This tendency is robust for all health states. This was our prior expectation.

[Insert table 3 about here]

Table 4 displays BCA confidence intervals at alpha = 0.05. It is apparent that differences between mean utilities are statistically significant by bootstrapping. This result holds regardless of means are trimmed. As same as before, differences hold for all health states.

[Insert table 4 about here]

Consistent with the bootstrap analysis, we also found significant differences between medians SG utilities (Wilcoxon Mann-Whitney test ; $p < 0.0001$). Differences are independent on the severity of the health status.

5. Discussion

This paper provides new evidence on how ‘irrelevant’ changes in the way we use to represent health risks can lead to inconsistent preferences. The novelty of this paper is that shows how superficially different frequency frames (e.g. X out of 100 vs Y out of 1,000) can distort standard gamble measurements. We find that utilities were significant higher when success and failure probabilities in the standard gamble were framed as frequencies with respect to 1,000 people in the population rather than as frequencies with respect to 100 people in the

population. This violation of invariance holds with independence on the severity of the health status and the statistical procedure used to test significance (bootstrap, Wilcoxon Mann-Whitney test).

The framing effect we have observed can be explained by the combined effect of a heuristic we name the numerator effect, and the phenomenon of loss aversion, formalized in Prospect Theory. The numerator effect predicts that the individual will focus on the numerator of a relative frequency, neglecting the denominator. This heuristic has been described in previous studies by Yamagishi (1997a) and Slovic et al. (2000). According to Prospect Theory outcomes are evaluated as losses or gains with respect to a reference point and loss aversion describes that people are more sensitive to losses than gains of equivalent size. The available evidence (Hershey and Schoemaker, 1985; Bleichrodt et al., 2001) suggests that in a standard gamble elicitation the sure outcome is taken as the reference point with respect to which both the best outcome (recovering full health) and the worst outcome (death) of the risky gamble are evaluated. In consequence, by loss aversion, the loss from the sure outcome to death with probability Y^* in 1,000 will be overweighted relative to the gain from the sure outcome to recovering full health with probability Y in 1,000, and, by the numerator effect, the absolute number of deaths Y^* would be regarded as excessive. Consequently, indifference between the sure outcome and the risky gamble will require that Y^* is lower than $10 \cdot X^*$, the numerator that would be necessary to be numerically equivalent relative frequencies $Y^*/1,000$ and $X^*/100$.

This paper is not without limitations. First, we did not test numeracy (i.e., facility with numbers) of participants in the experiment. It has been shown that being innumerate may distort perceptions of risk (Schwartz et al., 1997; Woloshin et al., 2000) even in case of highly educated samples as our case (Lipkus et al., 2001). Indeed, Woloshin et al. (2001) found that the validity of SG assessments is related to the subject's numerate. However, it should be noted that even though innumeracy can increase random error in probability judgments, it is not clear how it could yield a systematic bias as we have found in our experiment.

Second, the use of students as subjects may limit the generalizability of our findings. Hence it would be interesting to replicate the present study with general population. Nevertheless, many studies show that health

state valuations do not depend in a significant way on the representativeness of the study sample (see de Wit et al., 200 for a review).

Let us briefly explain the implications of our findings. It has been shown that the SG method is affected by loss aversion and probability weighting (Bleichrodt, 2001, 2002). It has been proposed quantitative corrections to these biases (Bleichrodt et al., 2001; van Osch et al., 2004; Bleichrodt et al., 2005). As we have just explained, our results are consistent with loss aversion and the numerator effect. Since both the numerator effect and probability weighting (i.e., nonlinear transformation of probabilities) are concerned with probabilities, perhaps it could be thought that invariance violations could be avoided by applying Prospect Theory corrective formulas. Unfortunately, that is not the case. Prospect Theory assumes that 10/1,000 is numerically equivalent to 1/100. Hence both probabilities are transformed in the same way (i.e., w(10/1,000) = w(1/100)). However, if 10/1,000 was perceived as higher than 1/100 because of the numerator effect, then to correct the bias 10/1,000 should be transformed in a different way from 1/100. Prospect Theory cannot manage this problem.

Although research has indicated that, in the presentation of information on risk to patients, relative frequency formats as we have used here are better than conventional probability formats (e.g., Bowling and Ebrahim, 2001), some authors (Jones et al, 1995; Yamagishi, 1997a,b; Schapira et al., 2004) have claimed that certain biases in probability judgments are more prevalent in frequency frames. For example, Yamagishi (1997b) found that upward and downward anchoring and subsequent biases occur in frequency judgments of social events. Schapira et al. (2001) found that a frequency format supported by human figures graphics was perceived to be of greater magnitude than a percentage-based format illustrated with a bar graph, when conveying a given numeric breast cancer risk. The current study finds another example that judgment in the frequency frame is not free from biases.

Efforts should be done in order to improve our knowledge on the heuristic of the numerator effect. Qualitative work (Schapira et la., 2001; Slovic et al., 2004) suggests that some people associate frequency formats with the risk of general population, whereas representations of risk in the form of percentages are associated with individual risks. According to this interpretation, probability of 10% of dead can be judged as an event (death) that

can occur or not. However, that 10 out of 100 people dead can create the affect-laden image of 10 deaths. The latter framing will be then judged as more dangerous than the former. Further research should be test whether new forms of presenting risk information can avoid the tendency to overweight the numerator of relative frequencies. For example, Yamagishi (1997a) suggests to compare the risk one wants to communicate as a comparison in probabilities with other risks following a descendent order. The question is if such a risk communication strategy could be used to elicit utilities with the standard gamble.

References

Armstrong K, Schwartz JS, Fitzgerald G, Putt M, Ubel P. Effect of framing as gain versus loss on understanding and hypothetical treatment choices: survival and mortality curves. Medical Decision Making 2002; 22: 76-83.

Banks SM, Salovey P, Greener S, Rothman AJ, Moyer A, Beauvais J, Epel E. The effects of message framing on mammography utilization. Health Psychology 1995; 14: 178-184.

Bernstein LM, Chapman GB, Elstein AS (1999). Framing effects in choices between multioutcome life-expectancy lotteries. Medical Decision Making 1999; 19: 324-338.

Bleichrodt H, Abellan-Perpiñan JM, Pinto JL, Mendez-Martinez I. Resolving inconsistencies in utility measurement under risk: tests of generalizations of expected utility. Working Paper UPF 2005; 718.

Bleichrodt H, Pinto J L, Wakker P P. Using Descriptive Findings of Prospect Theory to Improve the Prescriptive Use of Expected Utility. Management Science, 2001; 47: 1498-1514.

Bostic, R., Hernstein, R. J. and Luce, R. D. (1990). The effect on the preference reversal phenomenon of using choice indifferences. Journal of Economic Behavior and Organization 13, 193-212.

Bucher HC, Weinbacher M, Gyr K Influence of method of reporting study results on decision of physicians to prescribe drugs to lower cholesterol concentration, BMJ, 1994; 309: 761-4.

Denes-Raj V, Epstein S. Conflict between intuitive and rational processing: When people behave against their better judment. Journal of Personality and Social Psychology 1994; 66: 819-829.

Edwards AGK, Elwyn G, Mathews E., et al. Presenting risk information: a review of the effects of 'framing' and other manipulations on patient outcomes. Journal of Health Community 2001; 6: 61-82.

Eraker SA, Sox, HC. Assessment of patients' preferences for therapeutic outcomes. Medical Decision Making 1981; 1: 29-39.

Forrow L, Taylor WC, Arnold RM. Absolutely relative: how research results are summarized can affect treatment decisions. Am J Med. 1992; 92:121-4.

Gigerenzer G, Hoffrage U. How to improve Bayesian reasoning without instruction: Frequency formats. Psychology Review 1995; 102: 4-704.

Gurm HS, Litaker DG. Framing procedural risks to patients: is 99% safe the same as a risk of 1 in 100? Academic Medicine 2000; 75: 840-2.

Hersey JC, Schoemaker PJH. Probability versus certainty equivalence methods in utility measurement: Are they equivalent? Management Science, 1985; 31: 1213-1231.

Hoffrage U, Gigerenzer G. Using natural frequencies to improve diagnostic inferences. Academic Medicine 1995; 73: 538-540.

Hoffrage U, Lindsey S, Hertwig R, Gigerenzer G Communicating statistical information, Science. 2000;290:2261-2

Hux JE, Levinton CM, Naylor CD. Prescribing propensity: influence of life-expectancy gains and drug costs. JGIM 1994; 9: 195-201.

Hux JE, Naylor CD. Communicating the benefits of chronic preventive therapy:does the format of efficacy data determine patients' acceptance of treatment? Medical Decision Making 1995; 15: 152-7.

Kahneman D, Tversky A. Prospect theory: An analysis of decision under risk. Econometrica, 1979; 47: 263-291.

Kahneman, D. and Tversky, A. (1973) On the psychology of prediction, Psychology Review, 80, 237-251

Llewellyn-Thomas HA, McGreal J, Thiel EC. Cancer patients' decisión making and trial-entry preferences: the effects of "framing" information about short-term toxicity and long-term survival. Medical Decision Making 1995; 15:4-12.

Luce RD. Utility of gains and losses: Measurement-theoretical and experimental approaches. 2000. New Jersey: Lawrence Erlbaum Associates, Inc.

Malenka DJ, Baron JA, Johansen S, Wahrenberger JW, Ross JM. The framing effect of relative and absolute risk. Journal of General and Internal Medicine 1993; 8: 543-8.

McGettigan P, Sly K, O'Connell D, Hill S, Henry D. The effects of information framing on the practices of physicians. JGIM 1999; 14: 633-642.

McNeil B, Pauker S, Sox H, Tversky A. On the elicitation of preferences for alternative therapies, NEJM, 1982; 306: 1259-62.

O'Connor A. Effects of framing and level of probability on patients' preferences for cancer chemotherapy. Journal of Clinical Epidemiology 1989; 42:119-126.

O'Connor AM, Boyd NF, Tritchler DL, Kriukov Y, Sutherland H, Till JE. Eliciting preferences for alternative cancer drug treatments. The influence of framing, medium, and rater variables. Medical Decision Making 1985; 5:453-63.

Redelmeier DA, Rozin P, Kahneman D. Understanding patients' decisions: cognitive and emotional perspectives. JAMA 1993; 270: 72-6.

Sarfati D, Howden-Chapman P and Woodward, Salmon C Does the frame affect the picture? A study into how attitudes to screening for cancer are affected by the way benefit are expressed. Journal of Medical Screening, 1998; 5: 137-140.

Schapira MM, Nattinger AB, McHorney CA. Frequency or probability? A qualitative study of risk communication formats used in health care. Medical Decision Making 2001; 21: 459-467.

Schapira MM, Davids SL, McAuliffe TL, Nattinger AB. Agreement between scales in the measurement of breast cancer risk perceptions. Risk Analysis 2004; 24: 665-673.

Slovic P, Finucane ML, Peters E, MacGregor DG. The affect heuristic. In Gilovich T, Griffin D, Kahneman D (Eds.). Heuristic and biases: The Psychology of Intuitive Judgement 2002; 397-420. New York: Cambridge University Press.

Slovic P, Monahan J, MacGregor DG. Violence risk assessment and risk communication: the effects of using actual cases, providing instructions and employing probability versus frequency formats. Law and Human Behavior 2000; 24: 271-296.

Stalmeier PFM, Bezembinder TGG. The discrepancy between risky and riskless utilities: a matter of framing? Medical Decision Making 1999; 19: 435-47.

Tversky A, Kahneman D, Judgement under uncertainty: heuristics and biases, Science, 1974; 185: 453-458.

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

Tversky A, Kahneman D. The framing of decisions and the pshychology of choice. Science 1981; 211: 453-458.

Yamagishi K. When a 12.86% mortality is more dangerous than 24.14%: implications for risk communication. Applied Cognitive Psychology 1997a; 11: 495-506.

Hollander, M. and Wolfe, D. Nonparametric Statistical Methods, 1999, second edition, Wiley.

Efron, B. and Tibshirani, R.J., An Introduction to the Bootstrap, 1993, London: Chapman & Hall.

Barnett, V. and T. Lewis. Outliers in Statistical Data, 1995. Chichester: Wiley

Schwartz LM, Woloshin S, Black WC, Welch HG. The role of numeracy in understanding the benefit of screening mammography. AIM 1997; 127: 966-72.

Woloshin S, Schwartz LM, Moncur M, Gabriel S, Tosteson ANA. Assessing values for health: numeracy matters. Medical Decision Making 2001; 21: 382-90.

Lipkus I, Samsa G, Rimer BK. General Performance on a numeracy scale among highly educated samples. Medical Decision Making 2001: 21: 37-44.

De Wit, G. A., van Busschbach J. J., and de Charro F. T. (2000). Sensitivity and Perspective in the Valuation of Health Status. Health Economics, 9, 109-126.

Bleichrodt H. Probability weighting in choice under risk: an empirical test. Journal of Risk and Uncertaint 2001; 23, 185-198.

Bleichrodt, H. A New Explanation for the Difference Between Standard Gamble and Time Trade-Off Utilities. Health Economics 2002; 11, 447-456.

Van Osch S, Wakker PP, van den Hout WB, Stiggelbout AM. Correcting biases in standard gamble and time tradeoff utilities. Medical Decision Making 2004; 24: 511-517.

Bowling A, Ebrahim S. Measuring patients' preferences for treatment and perceptions of risk. Quality in Health Care 2001; 10 (Suppl I): i2-i8.

Jones SK, Jones KT, Frisch C. Biases of probability assessement: A comparison of frequency and single-case judgments. Organizational Behavior and Human Decision Processes 1995; 61: 109-122.

Yamagishi K. Upward versus downward anchoring in frequency judgements of social facts. Japonese Psychological Research 1997b; 2: 124-129.

Figures and tables

Table 1. The descriptions of the health states X, Y, Z, and W

Health state X	Health state Y
Some problems walking about Some problems with performing self care activities (e.g. eating, washing or dressing) No problems with performing usual activities (e.g. work, study, housework, family or leisure activities) No pain or discomfort Not anxious or depressed	No problems walking about No problems with performing self care activities (e.g. eating, washing or dressing) Some problems with performing usual activities (e.g. work, study, housework, family or leisure activities) Moderate pain or discomfort Moderately anxious or depressed
Health state Z	**Health state W**
Some problems walking about Some problems with performing self care activities (e.g. eating, washing or dressing) Some problems with performing usual activities (e.g. work, study, housework, family or leisure activities) Moderate pain or discomfort Moderately anxious or depressed	Some problems walking about Unable to wash or dress self Some problems with performing usual activities (e.g. work, study, housework, family or leisure activities) Moderate pain or discomfort Moderately anxious or depressed

Figure 1. Example of an elicitation with the choice-based procedure (1)

Figure 2. Example of an elicitation with the choice-based procedure (2)

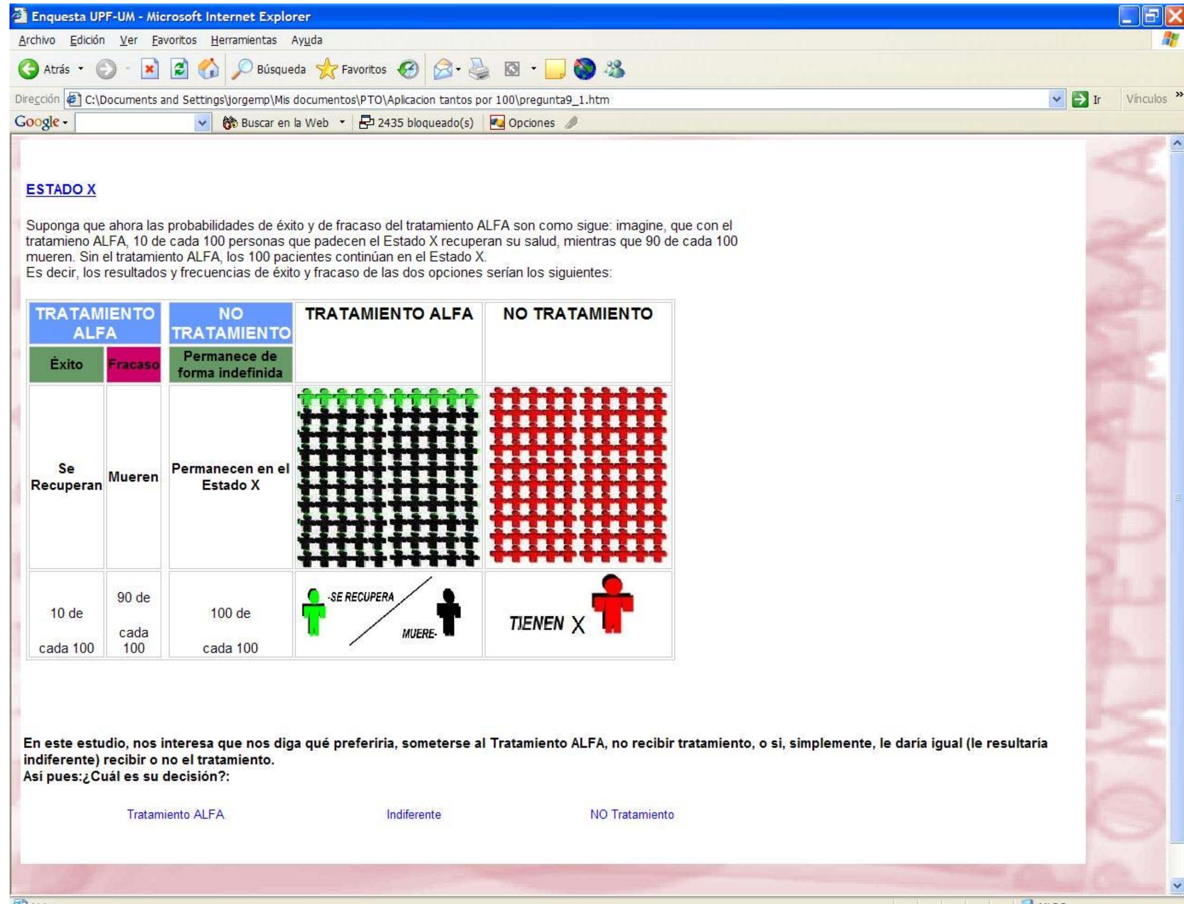

Table 2. Utilities elicited by means of the VAS and the SG (full sample, N=200). Means, medians and standard deviations (SD)

	Health state X		Health state W		Health state Z		Health state Y	
	VAS	SG	VAS	SG	VAS	SG	VAS	SG
Mean Median (SD)	0.592 0.600 (0.163)	0.846 0.900 (0.160)	0.550 0.550 (0.149)	0.838 0.900 (0.148)	0.303 0.300 (0.107)	0.624 0.625 (0.256)	0.166 0.150 (0.081)	0.437 0.450 (0.296)

Table 3. Utilities elicited by the SG (subsamples, N=100). Means, medians and standard deviation (SD)

	Health state X		Health state W		Health state Z		Health state Y	
	Out of 100	Out of 1,000	Out of 100	Out of 1,000	Out of 100	Out of 1,000	Out of 100	Out of 1,000
Mean Median (SD)	0.790 0.850 (0.185)	0.902 0.945 (0.105)	0.796 0.800 (0.161)	0.880 0.908 (0.120)	0.575 0.600 (0.271)	0.673 0.700 (0.232)	0.393 0.375 (0.272)	0.481 0.500 (0.313)

Table 4. Bootstrap BCA confidence intervals for the difference between mean SG utilities ($U(Q)_{SG1,000}$ - $U(Q)_{SG100}$).

Health state	Trimming	Eliminated values (Out of 1,000) left (right)	Eliminated values (Out of 100) left (right)	BCA confidence interval (α = 0.05)	
X	0	0(0)	0(0)	0.069	0.155
	0.5 σ	26(36)	20(40)	0.111	0.153
	σ	14(0)	15(15)	0.086	0.136
	1.5 σ	8(0)	11(0)	0.055	0.113
	2 σ	3(0)	8(0)	0.052	0.118
	2.5 σ	2(0)	1(0)	0.071	0.149
W	0	0(0)	0(0)	0.046	0.125
	0.5 σ	24(41)	29(42)	0.080	0.116
	σ	18(0)	15(16)	0.087	0.141
	1.5 σ	15(0)	9(0)	0.065	0.120
	2 σ	2(0)	4(0)	0.043	0.112
	2.5 σ	2(0)	2(0)	0.048	0.121
Z	0	0(0)	0(0)	0.028	0.163
	0.5 σ	30(35)	30(35)	0.074	0.139
	σ	15(19)	22(21)	0.037	0.129
	1.5 σ	6(0)	10(3)	0.030	0.150
	2 σ	6(0)	0(0)	0.066	0.200
Y	0	0(0)	0(0)	0.017	0.169
	0.5 σ	37(36)	41(30)	0.015	0.099
	σ	24(24)	26(20)	0.019	0.135
	1.5 σ	2(5)	0(6)	0.024	0.183
	2 σ	0(0)	0(4)	0.036	0.186