A Parameter-Free Elicitation of the Probability Weighting Function in Medical Decision Analysis

Han Bleichrodt • Jose Luis Pinto
iMTA/iBMG, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
Department of Economics, Universitat Pompeu Fabra, Barcelona, Spain
bleichrodt@bmg.eur.nl • jose.pinto@econ.upf.es

An important reason why people violate expected utility theory is probability weighting. Previous studies on the probability weighting function typically assume a specific parametric form, exclude heterogeneity in individual preferences, and exclusively consider monetary decision making. This study presents a method to elicit the probability weighting function in rank-dependent expected utility theory that makes no prior assumptions about the functional form of the probability weighting function. We use both aggregate and individual subject data, thereby allowing for heterogeneity of individual preferences, and we examine probability weighting in a new domain, medical decision making. There is significant evidence of probability weighting both at the aggregate and at the individual subject level. The modal probability weighting function is inverse S-shaped, displaying both lower subadditivity and upper subadditivity. Probability weighting is in particular relevant at the boundaries of the unit interval. Compared to studies involving monetary outcomes, we generally find more elevation of the probability weighting function. The robustness of the empirical findings on probability weighting indicates its importance. Ignoring probability weighting in modeling decision under risk and in utility measurement is likely to lead to descriptively invalid theories and distorted elicitations.
(*Nonexpected Utility*; *Decision Theory*; *Probability Weighting*; *Utility Assessment*; *Medical Decision Making*)

It is by now widely acknowledged that expected utility theory is not valid as a descriptive theory of choice under risk. An important reason why people violate expected utility theory is that their preferences between risky prospects are not linear in probabilities. In response to the observed violations of expected utility theory, several nonexpected utility theories have been proposed. The most important theories among these nonexpected utility theories are rank-dependent expected utility theory (Quiggin 1981, Yaari 1987) and its derivative cumulative prospect theory (Starmer and Sugden 1989, Luce and Fishburn 1991, Tversky and Kahneman 1992). An essential characteristic of the latter two theories is that probabilities do not enter linearly in the evaluation formula, but are transformed into decision weights through a cumulative probability weighting function. The nonlinearity of preferences in probability makes it possible to explain choice patterns that are at variance with expected utility theory. A disadvantage of using transformed probabilities in the evaluation formula is that the elicitation or estimation of the model becomes more involved, because in

0025-1909/00/4611/1485$05.00
1526-5501 electronic ISSN

Management Science © 2000 INFORMS
Vol. 46, No. 11, November 2000 pp. 1485–1496

addition to the utility function, the probability weighting function has to be elicited.

Previous studies that elicited the probability weighting function in rank-dependent expected utility theory or cumulative prospect theory generally have three characteristics in common. First, they estimated both the utility function and the probability weighting function by parametric techniques (Tversky and Kahneman 1992, Camerer and Ho 1994, Tversky and Fox 1995, Wu and Gonzalez 1996). Specific inverse S-shaped functional forms were suggested for the probability weighting function (Lattimore et al. 1992, Tversky and Kahneman 1992, Prelec 1998). A disadvantage of this approach is that the estimations depend critically on the assumed functional form. If the true functional form is different from the assumed functional form, then conclusions drawn from the estimations need no longer hold. Several studies (Currim and Sarin 1989, Wu and Gonzalez 1996) used nonparametric methods to derive qualitative properties of the probability weighting function. However, these nonparametric techniques were not used to estimate quantitative probability weights.

Second, these studies have been based on aggregate data: They either used a single-agent stochastic choice model or fitted the weighting function to the median subject. Thereby, heterogeneity of individual preferences is ruled out.

Third, they focused on one specific outcome domain: money. Little is known about the generalization of their findings to other decision domains. Previous research suggests that probability weighting may depend on the decision context. Currim and Sarin (1989), for instance, argue that the outcome level may affect probability weighting. Several studies report evidence that the shape of the probability weighting function in decision under uncertainty depends on the source of the uncertainty (Heath and Tversky 1991, Tversky and Fox 1995, Kilka and Weber 1998). Similarly, the outcome domain can affect probability weighting. Wakker and Deneffe (1996) found higher risk aversion for life duration than for money even though utility curvature was similar for these outcomes. Under rank-dependent expected utility theory, this can only be explained by a difference in probability weighting. Recently, Rottenstreich and Hsee (1999) have presented evidence that probability weighting depends on the outcome domain.

This article generalizes the aforementioned studies on the probability weighting function in rank-dependent expected utility theory in three respects. First, we elicit the probability weighting function without making any prior assumptions about its functional form. That is, we provide a parameter-free elicitation of the probability weighting function. Second, we use both aggregate data and individual subject data. Third, we examine probability weighting in a new domain: medical decision making.

We apply the trade-off method of Wakker and Deneffe (1996) to elicit first the utility function and then the probability weighting function, using the elicited utilities as inputs. Independently from us, two other papers have also provided parameter-free assessments of the probability weighting function using both aggregate data and individual subject data. Gonzalez and Wu (1999) use an alternating least squares approach to estimate the probability weighting function and the utility function simultaneously. Similar to our study, Abdellaoui (2000) applies the trade-off method to elicit the utility function first and then the probability weighting function, using the elicited utilities as inputs. Even though he also uses the trade-off method, Abdellaoui's procedure is different from ours. We compare our approach with Abdellaoui's approach in §2. Both Gonzalez and Wu and Abdellaoui use monetary outcomes.

In what follows, §1 reviews rank-dependent expected utility theory and empirical evidence on probability weighting. Section 2 describes the trade-off method and our procedure to elicit the probability weighting function. Section 3 describes the experimental procedures used to elicit utilities and probability weights. The results are described in §4. Section 5 concludes.

1. Rank-Dependent Expected Utility Theory

Let X be a set of *outcomes*, in our case, life duration. We study decision under risk, and therefore assume a set of simple probability distributions P defined over X.

A typical element of P is the lottery $[p_1, x_1; \ldots; p_m, x_m]$, which yields outcome $x_i \in X$ with probability p_i. Here m is a positive integer and $p_1 + \cdots + p_m = 1$. Two lotteries $[p_1, x_1; \ldots; p_m, x_m]$ and $[p_1, y_1; \ldots; p_m, y_m]$ that induce the same ranking of outcomes are *comonotonic*. Let $[p_1, x_1; \ldots; p_m, x_m]$ be a lottery for which $x_1 \succeq \cdots \succeq x_m$, where $\succeq$ stands for ''at least as good as''. The rank-dependent expected utility of this lottery is equal to

$$RDEU[p_1, x_1; \ldots; p_m, x_m] = \sum_{i=1}^{m} \pi_i u(x_i), \tag{1}$$

where $\pi_i = w(\sum_{j=1}^{i} p_i) - w(\sum_{j=1}^{i-1} p_i)$, in particular $\pi_1 = w(p_1)$. The *probability weighting function* w is a strictly increasing function from [0, 1] to [0, 1] with $w(0) = 0$ and $w(1) = 1$. If w is the identity function, rank-dependent expected utility theory is identical to expected utility theory.

Several studies have provided preference conditions allowing different shapes of the probability weighting function (Tversky and Wakker 1995, Wu and Gonzalez 1996, Wu and Gonzalez 1998, Prelec 1998). Tversky and Wakker define two conditions to characterize the probability weighting function: lower subadditivity and upper subadditivity. *Lower subadditivity* means that a lower interval $[0, q]$ has more impact on a decision maker than an intermediate interval $[p, p + q]$, provided that $p + q$ is bounded away from one. Alternatively stated, lower subadditivity says that a change from impossible to possible has a stronger impact on an individual's decision than an equal change from possible to more possible. This effect is referred to as the *possibility effect*. *Upper subadditivity* says that an upper interval $[1 - q, 1]$ has more impact than an intermediate interval $[p, p+q]$, provided that p is bounded away from zero. Hence, a change from possible to certain has more impact than an equal change from possible to more possible. This effect is referred to as the *certainty effect*. The effect of lower subadditivity and upper subadditivity is to produce an inverse S-shaped probability weighting function, overweighting small probabilities and underweighting intermediate and high probabilities. In the context of rank-dependent expected utility, the probability weighting function satisfies lower subadditivity if $w(q) \geq w(p + q) - w(p)$, provided $w(p + q)$ is bounded away from one, and it satisfies upper subadditivity if $1 - w(q) \geq w(p+q) - w(p)$, provided $w(p)$ is bounded away from zero.

Several parametric specifications of the probability weighting function have been suggested in the literature. Tversky and Kahneman (1992) proposed the following one-parameter specification:

$$w(p) = \frac{p^\gamma}{[p^\gamma + (1 - p)^\gamma]^{\frac{1}{\gamma}}}. \tag{2}$$

This function is monotonic and has an inverse S-shape for values of γ between 0.27 and 1.

Gonzalez and Wu (1999) suggest a two-parameter specification for the inverse S-shaped probability weighting function, adopted before by Goldstein and Einhorn (1987), Lattimore et al. (1992), and Tversky and Fox (1995):

$$w(p) = \frac{\delta p^\gamma}{\delta p^\gamma + (1 - p)^\gamma}. \tag{3}$$

In Equation (3), the parameter γ primarily controls curvature of the probability weighting function, i.e., the extent to which people are able to discriminate between differences in probability, and the parameter δ primarily controls elevation, i.e., the extent to which people find the chance domain attractive.

Prelec (1998) has axiomatized alternative specifications for the inverse S-shaped probability weighting function. His proposed one-parameter specification is

$$w(p) = \exp(-(-\ln p)^\alpha), \tag{4}$$

and his two-parameter specification is

$$w(p) = \exp(-\beta(-\ln p)^\alpha). \tag{5}$$

The interpretation of the parameters α and β is similar to Equation (3): α primarily controls curvature and β primarily controls elevation.

Table 1 summarizes the results of some empirical studies that estimated the parameters in the above specifications. Cumulative prospect theory allows the parameters to be different for gains and losses. All estimates are consistent with an inverse S-shaped weighting function.

Table 1 Empirical Studies on the Probability Weighting Function

Functional Form	Parameter Estimates
$w(p)=\dfrac{p^{\gamma}}{[p^{\gamma}+(1-p)^{\gamma}]^{\frac{1}{\gamma}}}$	Tversky and Kahneman (1992): $\gamma=0.61$ (gains), $\gamma=0.69$ (losses)
	Camerer and Ho (1994): $\gamma=0.56$ (gains)
	Wu and Gonzalez (1996): $\gamma=0.71$ (gains)
	Abdellaoui (2000): $\gamma=0.60$ (gains), $\gamma=0.70$ (losses)
$w(p)=\dfrac{\delta p^{\gamma}}{\delta p^{\gamma}+(1-p)^{\gamma}}$	Wu and Gonzalez (1996): $\delta=0.84$, $\gamma=0.68$ (gains)
	Gonzalez and Wu (1999): $\delta=0.77$, $\gamma=0.44$ (gains)
	Tversky and Fox (1995): $\delta=0.77$, $\gamma=0.69$ (gains)
	Abdellaoui (2000): $\delta=0.65$, $\gamma=0.60$ (gains)
	Abdellaoui (2000): $\delta=0.84$, $\gamma=0.65$ (losses)
$w(p)=\exp(-(-\ln p)^{\alpha})$	Wu and Gonzalez (1996): $\alpha=0.74$ (gains)

2. The Trade-Off Method

The elicitation procedure consisted of two parts. In the first part, we elicited the utility function for life duration. The elicited utilities were then used as inputs in the second part to elicit the probability weights. We used the trade-off method for the elicitation of utilities and probability weights (Wakker and Deneffe 1996). This method was selected because it is not affected by probability weighting. Alternative techniques such as the probability equivalence method, the certainty equivalence method, and the lottery equivalence method (McCord and de Neufville 1986) suffer from the defect that they are vulnerable to probability weighting and do not provide valid utilities under rank-dependent expected utility.

Part 1: Elicitation of the Utility Function

The trade-off method determines a standard sequence of outcomes, which are equally spaced in terms of utility. The first step consists of the selection of two reference outcomes R and r with $R \succeq r$ and a starting outcome x_0. Then an individual is asked to specify x_1 such that he is indifferent between the lotteries $[p, R; 1-p, x_0]$ and $[p, r; 1-p, x_1]$, with $R \succeq x_0$ and $r \succeq x_1$ to ensure that the two lotteries are comonotonic. After x_1 has been elicited, the individual is asked to specify the number x_2 such that he is indifferent between $[p, R; 1-p, x_1]$ and $[p, r; 1-p, x_2]$. If $r \succeq x_2$ and rank-dependent expected utility theory holds, then the first indifference yields

$$\begin{aligned}
&[p, R; 1-p, x_0] \sim [p, r; 1-p, x_1] \\
&\Leftrightarrow w(p)U(R) + [1-w(p)]U(x_0) \\
&= w(p)U(r) + [1-w(p)]U(x_1) \\
&\Leftrightarrow w(p)[U(R)-U(r)] \\
&= [1-w(p)][U(x_1)-U(x_0)], \qquad (6a)
\end{aligned}$$

and, similarly, the second indifference yields

$$\begin{aligned}
&[p, R; 1-p, x_1] \sim [p, r; 1-p, x_2] \\
&\Leftrightarrow w(p)[U(R)-U(r)] = [1-w(p)][U(x_2)-U(x_1)]. \qquad (6b)
\end{aligned}$$

Combining (6a) and (6b) gives

$$U(x_2) - U(x_1) = U(x_1) - U(x_0). \qquad (7)$$

As long as $r \succeq x_j$, we can proceed to ask for indifference between $[p, R; 1-p, x_{j-1}]$ and $[p, r; 1-p, x_j]$, in the process eliciting a standard sequence $(x_0, \ldots, x_k)$ for which $U(x_i)-U(x_{i-1}) = U(x_j)-U(x_{j-1})$ for all $1 \le i, j \le k$. Given the uniqueness properties of the utility function U, the scale and the origin of the function can be chosen arbitrarily. We used the scaling $U(x_0)=0$ and $U(x_k)=1$, from which it follows that for all $1 \le j \le k: U(x_j)=j/k$.

Part 2: Elicitation of the Probability Weighting Function

The probability p was held constant throughout the elicitation of the standard sequence. This ensured that the elicited utilities were not distorted by probability weighting. Equations (6a) and (6b) show that the terms $w(p)$ and $1-w(p)$ cancel out if p is held constant. To elicit the probability weighting function, the probabilities have to be varied. Probability weights were determined by two types of questions. For low probabilities, we asked for an outcome z_r such that the individual is indifferent between $[p', x_i; 1-p', x_j]$ and $[p', x_k; 1-p', z_r]$ with $x_k \ge x_i \ge x_j$, and x_i, x_j, and x_k elements of the standard sequence elicited in the first part. For higher probabilities, we asked for an outcome z_s

such that indifference holds between $[p', x_m; 1-p', x_n]$ and $[p', z_s; 1-p', x_q]$ with $x_m \geq x_n \geq x_q$ and x_m, x_n, and x_q elements of the standard sequence. We explain below why we used different questions for low and high probabilities.

By rank-dependent expected utility, the weight of probability p' is determined from the first indifference as

$$w(p') = \frac{u(x_j) - u(z_r)}{[u(x_j) - u(z_r)] + [u(x_k) - u(x_i)]} \tag{8}$$

and from the second indifference as

$$w(p') = \frac{u(x_n) - u(x_q)}{[u(z_s) - u(x_m)] + [u(x_n) - u(x_q)]}. \tag{9}$$

As mentioned in the introduction, Abdellaoui (2000) also used the trade-off method to elicit first the utility function and then the probability weighting function with the elicited utilities as inputs. His procedure consists of the selection of an element x_j from the elicited standard sequence and to determine the probability p' that makes the individual indifferent between the lotteries $[p', x_k; 1-p', x_0]$ and $[1, x_j]$, where $[1, x_j]$ stands for x_j with certainty. Under rank-dependent expected utility theory and the chosen scaling, this indifference determines the weight of probability p' as

$$w(p') = \frac{j}{k}. \tag{10}$$

If p' is determined by a probability matching question, then the response scale differs between the two parts of the elicitation procedure: In the first part the outcome dimension is used to elicit indifference, in the second part the probability dimension. Previous studies have shown that different response scales prime different aspects of the decision problem, a phenomenon referred to as *scale compatibility* (Tversky et al. 1988). Scale compatibility has been observed both in matching and in sequential choice tasks (Delquié 1993, 1997). Abdellaoui (2000) makes a careful attempt to avoid the distorting impact of scale compatibility by eliciting indifference through nonsequential choice questions. No experimental evidence exists about the impact of scale compatibility in nonsequential choice tasks.

To avoid the distorting impact of changing response scales, we used only the outcome dimension to elicit indifferences. Our procedure has three potential disadvantages. First, the outcomes z_r and z_s in Equations (8) and (9) need not belong to the standard sequence, in which case their utility has to be estimated from the utility of elements of the standard sequence. This approximation may introduce bias. However, the utility function does not deviate strongly from linearity over small intervals (Wakker and Deneffe 1996), and a linear approximation will be reasonable as long as the standard sequence is sufficiently fine.

Second, our procedure imposes bounds on the elicited probability weights. Because z_r can never be less than zero, Equation (8) forces the probability weights to lie between zero and $u(x_j)/([u(x_j)] + [u(x_k) - u(x_i)])$. Hence, the elicited probability weights are bounded above and, therefore, we only used Equation (8) to elicit the weights of lower probabilities. The utility of z_s in Equation (9) only be determined if z_s is smaller than x_k, the final element in the standard sequence. Therefore application of Equation (9) leads to probability weights that lie between $(u(x_n) - u(x_q))/([1 - u(x_m)] + [u(x_n) - u(x_q)])$ and one. That is, the elicited probability weights are bounded below, and we therefore only used Equation (9) to elicit the weights of higher probabilities. The ''boundedness problem'' can be limited by an appropriate choice of the elements of the standard sequence. If x_j in Equation (8) is relatively far from x_0 in the standard sequence and x_k and x_i are relatively close, then the elicited probability weight can take on values close to one. Similarly, if in Equation (9) x_n is relatively close to x_q and x_m is relatively far from x_k, then the elicited probability weight can take on values close to zero. In §3 we explain how we have handled the ''boundedness problem'' in our experimental design, and in §4 we present evidence that it caused no problems in our data.

Third, our method may suffer from error propagation. Equations (8) and (9) determine probability weights by a ratio. Error propagation for ratios can be problematic if the denominator is close to zero, so that small errors in the numerator induce large errors in the quotient. Such problems do not occur in Equations (8) and (9) because the denominator is remote from zero, more than the numerator. Additionally, the numerator and the denominator are positively correlated because of the common term, which again reduces the overall error in the quotient. These analytical observations suggest that error propagation will not be dramatic in our design. To obtain insight into the extent to which our procedure is affected by error propagation, we performed two simulation studies based on two different error theories. These studies are described in §4. Their results indicate that error propagation is not a problem in our study.

3. Experiment

3.1. Subjects

Prior to the main experiment, the questionnaire was tested in several pilot sessions, using university staff as subjects. Fifty-one subjects participated in the main experiment. All subjects were undergraduate economics students from the University of Pompeu Fabra. The subjects were paid 5,000 Pesetas (approximately 30 U.S. dollars) for their participation. Because we used life duration as the outcome domain, individual responses to the experimental questions could not be played out for real. That is, there were no real incentives in our study. Several studies have argued and presented empirical evidence that hypothetical and real questions give similar results in decision under risk (Tversky and Kahneman 1992, Beattie and Loomes 1997). In a recent review of the literature on the effect of financial incentives in experiments, Camerer and Hogarth (1999) conclude that incentives appear to help most frequently in judgment and decision tasks that are different from the task we used. The trade-off method assesses preferences between lotteries, and for such a task real incentives do not seem to improve performance.

3.2. Procedures

The experiment was carried out in two personal interview sessions separated by two weeks. Personal interview sessions were used in an attempt to obtain high-quality data.

Both sessions started with an explanation of the trade-off method, both orally and in writing. Subjects were told that they suffered from one of two diseases, but that it was right now unknown from which disease they suffered. The diseases were anonymously labeled A and B to avoid possible framing effects. Subjects were further informed that it is known from previous medical experience that people with the symptoms they displayed have Disease A half of the time and Disease B half of the time. There exist two treatments to beat the symptoms, but the effectiveness of the treatments depends on the disease. The outcomes of the treatments were numbers of remaining life duration. Subjects were told that the remaining life duration was spent in good health. Subjects had to choose a treatment before it was known which disease they actually had.

Following the explanation of the decision problem, the subjects were given a practice question and asked to explain their answer. Their explanation indicated whether they understood the questions and the experimental task. After we were convinced that a subject understood the questions, we moved on to the actual experiment.

The first experimental session started with the determination of the standard sequence for utility. The reference outcomes R and r were set at 55 years and 45 years, respectively. We had learned from the pilot sessions that these reference values created a standard sequence in which the elements were fairly close together (1 to 5 years). Outcome x_0 was set equal to zero, that is, x_0 corresponded to immediate death. A standard sequence $x_1, \ldots, x_6$ was elicited by determining the number of years x_j, $j=1, \ldots, 6$ for which subjects were indifferent between $[\frac{1}{2}, 55; \frac{1}{2}, x_{j-1}]$ and $[\frac{1}{2}, 45; \frac{1}{2}, x_j]$. The elicited standard sequence traces the utility function for life duration. The utility function was scaled such that $U(x_j)=j/6$.

The pilot sessions had shown that people find the trade-off method easier to answer if they first determine the life durations for which one of the two treatments is clearly superior and then move towards the indifference value. We therefore asked subjects first to compare the treatments $[\frac{1}{2}, 55; \frac{1}{2}, x_{j-1}]$ and $[\frac{1}{2}, 45; \frac{1}{2}, x_j]$ for $x_j=x_{j-1}$ and for $x_j=45$. All subjects agreed that the treatment $[\frac{1}{2}, 55; \frac{1}{2}, x_{j-1}]$ is better than the treatment $[\frac{1}{2}, 45; \frac{1}{2}, x_{j-1}]$, and all but one subject agreed that the treatment [1, 45] is better than the treatment

$[\frac{1}{2}, 55; \frac{1}{2}, x_{j-1}]$.[1] Subjects were then told that these preferences imply that there should be a value of x_j between x_{j-1} and 45 for which their preferences between the treatments switch. They were asked to determine this ''switching value'' by gradually increasing x_j starting from x_{j-1} and by gradually decreasing x_j from 45 years until they arrived at a range of values for which they found it hard to choose between the treatments. From this range of values, subjects were then asked to pick the value of x_j for which they considered the treatments most finely balanced. This procedure is similar to that of Dubourg et al. (1994).

Over the two sessions, five questions were asked to elicit the shape of the probability weighting function. Weights were established for the following five probabilities: 0.10, 0.25, 0.50, 0.75, and 0.90. These probabilities were selected to include both probabilities that are typically overweighted (0.10 and 0.25) and probabilities that are typically underweighted (0.50, 0.75, and 0.90) according to previous research.

The weights for probabilities 0.10, 0.25, and 0.50 were elicited by asking for the indifference value z_j in the comparison between $[p, x_4; 1-p, x_3]$ and $[p, x_5; 1-p, z_j]$ and by applying Equation (8). This question leads to an upper bound of the probability weight of 0.75. This upper bound is reached if $z_j = 0$. We denote the responses to the questions for $p = 0.10$, 0.25, and 0.50 by z_1, z_2, and z_3, respectively.

The weights for probabilities 0.75 and 0.90 were elicited by asking for the indifference value z_j in the comparison between $[p, x_3; 1-p, x_2]$ and $[p, z_j; 1-p, x_1]$ and by applying Equation (9). This question leads to a lower bound of the probability weight of 0.25. This lower bound is reached if $z_j = x_6$. We denote the responses to the questions for $p = 0.75$ and 0.90 by z_4 and z_5, respectively.

The procedure used to elicit z_1 through z_5 was similar to the procedure used in the elicitation of the utility function. Subjects were encouraged to determine first the values of z_j, $j = 1, \ldots, 5$, for which they clearly preferred one of the treatments, and finally the value of z_j for which they considered the two treatments most finely balanced. The utilities of z_1 through z_5 were determined both under the assumption that utility is linear between points of the standard sequence, the linear approximation, and under the assumption that utility is a power function, $U(x) = \mu x^{\theta}$. A power function was selected because this function is frequently used in the literature (e.g., Tversky and Kahneman 1992) and there exists empirical support for a power function for life duration (Pliskin et al. 1980, Stiggelbout et al. 1994). The power function was estimated using the elements of the standard sequence and their corresponding utilities as data inputs.

In the first experimental session we asked for z_1, z_3, and z_4. The order of the questions was varied to avoid order effects. The second session served to elicit z_2 and z_5 and to repeat three questions from the first session to test the consistency of subjects' answers. The questions that were repeated varied across subjects. Procedures and methods in the second session were identical to those in the first session.

[1] This subject preferred $[\frac{1}{2}, 55; \frac{1}{2}, x_4]$ to $[1, 45]$. His preferences violate comonotonicity of the lotteries, and he was excluded from the analyses for this reason.

4. Results

4.1. Reliability and Consistency

Two subjects were excluded from the analyses: one was unable to make any trade-offs; the other subject's responses violated comonotonicity—both x_5 and x_6 exceeded r for this subject (see also Footnote 1).

The second session responses slightly exceeded the first session responses. The mean difference between first and second session responses was -0.327 ($SE = 0.184$). The difference between first and second session responses did not reach conventional levels of significance by a paired t test. ($t_{109} = -1.778$, $p = 0.078$).

4.2. Utility Curvature

Aggregate Data. Figure 1 displays the elicited utility function for life duration. The difference between successive points of the standard sequence (the step size), increases, leading to a concave utility function. The step size increases gradually and, hence, the linear utility function for life duration is a good approximation over short intervals.

Figure 1 The Elicited Utility Function for Life Duration

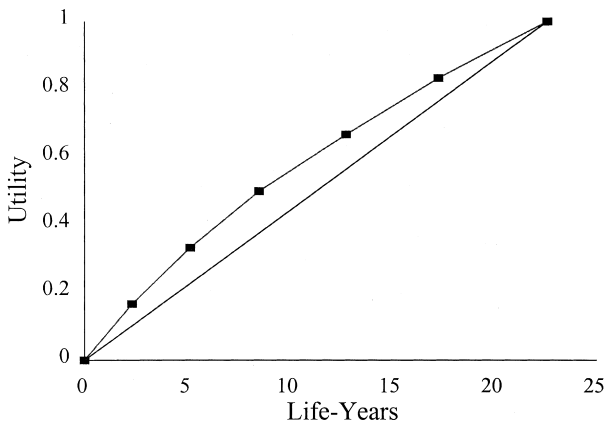
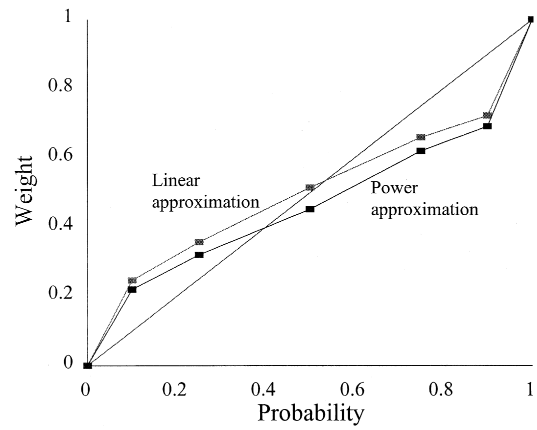


Figure 2 The Elicited Probability Weighting Function



Note. The median probability weights are 0.253, 0.357, 0.526, 0.668, and 0.707 under the linear approximation and 0.224, 0.320, 0.462, 0.630, and 0.677 under the power approximation.

Individual Data. By Δ^j_{j-1} we denote the difference $(x_j - x_{j-1}) - (x_{j-1} - x_{j-2})$, that is, the difference between two successive step sizes of the standard sequence. Positive Δ^j_{j-1} corresponds to a concave utility function, zero Δ^j_{j-1} to a linear utility function, and negative Δ^j_{j-1} to a convex utility function. For each subject, we observe five values of Δ^j_{j-1}. There are 29 (9) subjects with at least 3 (4) positive values of Δ^j_{j-1}, 13 (5) subjects with at least 3 (4) zero values of Δ^j_{j-1}, and 1 (0) subject with at least 3 (4) negative values of Δ^j_{j-1}. Clearly, the modal shape of the utility function is concave.

4.3. Probability Weighting

Boundedness Problem and Violations of Stochastic Dominance. No subject reported either a value of zero in the questions involving z_1, z_2, and z_3, or a value of x_6 in the questions involving z_4 and z_5. We conclude that the boundedness of the probability weights caused no problems in our data.

Some responses were such that the outcomes of one treatment were better under both states of the world and, hence, this treatment stochastically dominated the other treatment. Because it is not plausible that an individual is indifferent between two treatments where one treatment stochastically dominates the other, these responses were interpreted as reflecting confusion and were excluded from the analyses.

Estimation of the Power Utility Function. A power function was estimated for each subject based on the criterion ''minimize the sum of squared residuals''. This function was used in the power approximation. The mean of the individual optimal estimates of the power coefficient θ is 0.779 (median $=0.769$; $SE=0.0177$), which is close to 0.74, the estimate obtained by Stiggelbout et al. (1994) in a group of cancer patients. The estimated utility functions fit the data very well: The overall proportion of the total variation explained by the power function (R^2) is equal to 0.987.

4.3.1. Aggregate Data. Figure 2 displays the mean probability weights under the linear and the power approximation. The pattern of probability weights is consistent with an inverse S-shaped probability weighting function: Small probabilities are overweighted and intermediate and large probabilities are underweighted. The only deviation from the inverse S-shape is that the slope of the weighting function between 0.50 and 0.75 exceeds the slope between 0.75 and 0.90. The degree of upper subadditivity exceeds the degree of lower subadditivity. Compared to both the parametric and the nonparametric studies using monetary outcomes, we find more lower subadditivity and similar upper subadditivity. The exception is the nonparametric study by Gonzalez and Wu (1999), who find lower and upper subadditivity similar to ours.

The probability weights under the power approximation are smaller than the weights under the linear approximation. This happens because most subjects have a concave utility function for life duration. If utility is concave, then the linear approximation will underestimate the utility of z_r and z_s. It can be verified from Equations (8) and (9) that the probability weight is negatively related to the utility of z_r, respectively z_s. Hence, if utility is concave the linear approximation tends to overestimate the probability weight.

The note under Figure 2 shows the median data. The use of median instead of mean data does not affect the conclusions.

4.3.2. Individual Data. Let ∂^j_{j-1} denote the average slope of the probability weighting function between probabilities j and $j-1$:

$$\partial^j_{j-1} = \frac{w(p_j) - w(p_{j-1})}{p_j - p_{j-1}}.$$

Let ∇^j_{j-1} denote the change in the average slope between successive probabilities, i.e., the difference between ∂^j_{j-1} and ∂^{j-1}_{j-2}. The ∇^j_{j-1} can be used to examine the shape of the probability weighting function. For example, concavity of the probability weighting function corresponds to a decreasing slope and, hence, to negative ∇^j_{j-1}. Similarly, linearity corresponds to zero ∇^j_{j-1} and convexity to positive ∇^j_{j-1}. An inverse S-shaped probability weighting function is concave for small j, and convex for larger j.

For each subject, we observed five values of ∇^j_{j-1}. A subject's probability weighting function was classified as lower subadditive if $\nabla^{0.25}_{0.1}$ was negative. The probability weighting function was upper subadditive if $\nabla^1_{0.9}$ was positive. The probability weighting function was concave if at least three ∇^j_{j-1} were negative and the subject did not exhibit upper subadditivity, linear if at least three ∇^j_{j-1} were zero and the subject did not display both upper and lower subadditivity, and convex if at least three ∇^j_{j-1} were positive and the subject did not exhibit lower subadditivity.

Table 2 shows the classification of individuals according to the shape of their probability weighting function. The table shows strong evidence for lower and upper subadditivity. A small minority of subjects has a concave or convex weighting function and no subject has a linear weighting function.

Table 2 **Classification of Subjects According to the Slope of Their Probability Weighting Function**

Shape	Proportion of Subjects (Linear Approximation)	Proportion of Subjects (Power Approximation)
Concave	10.9%	6.5%
Linear	0%	0%
Convex	4.4%	8.7%
Lower Subadditivity	95.7%	91.3%
Upper Subadditivity	86.9%	89.1%
Lower and Upper Subadditivity	83.7%	81.4%

Table 3 **Classification of Subjects Based on $w(0.10)$ and $1 - w(0.90)$**

Shape	Proportion of Subjects (Linear Approximation)	Proportion of Subjects (Power Approximation)
Lower Subadditivity	95.8%	91.7%
Upper Subadditivity	91.3%	95.7%
Lower and Upper Subadditivity	88.9%	88.9%
PE Exceeds CE	47.5%	25.0%
PE Equal to CE	10.0%	0%
CE Exceeds PE	42.5%	75.0%

Note. PE stands for ''probability effect'' and CE for ''certainty effect.''

An alternative way to examine lower subadditivity and upper subadditivity at the individual subject level is to look at $w(0.10)$ and $1 - w(0.90)$. A subject satisfies lower subadditivity if $w(0.10) > 0.10$, and upper subadditivity if $1 - w(0.90) > 0.10$. This test is comparable to the metric used by Tversky and Fox (1995). This metric also permits a test of the relative sizes of the possibility effect and the certainty effect by comparing $w(0.10)$ to $1 - w(0.90)$. The possibility effect exceeds the certainty effect if $w(0.10) > 1 - w(0.90)$.

Table 3 displays the analysis based on the above metric. The table confirms both lower and upper subadditivity. Among the subjects who satisfy both lower and upper subadditivity, there are slightly more subjects for whom the possibility effect exceeds the certainty effect than subjects for whom the certainty effect exceeds the possibility effect under the linear approximation. Under the power approximation this conclusion is reversed. This reversal in conclusion occurs because for many subjects the possibility effect and the certainty effect are approximately equal.

Linearity of w for Intermediate Probabilities. There exists some controversy in the literature about the question of whether the probability weighting function is nonlinear throughout the unit interval [0,1] or whether nonlinearities occur only at the boundaries of the unit interval (Camerer 1992, Wu and Gonzalez 1996). In the latter case, the probability weighting function is linear, and hence consistent with expected utility theory, for intermediate probabilities. Wu and Gonzalez (1996) found support for nonlinearity, with weights becoming less concave throughout the unit interval, whereas Camerer's (1992) results support linearity away from the boundaries. Abdellaoui (2000) obtains mixed results.

We tested for linearity of the probability weighting function by examining the ∂_{j-1}^{j} for $j = 0.25$, 0.50, 0.75, 0.90. If the probability weighting function is linear in the interior of the unit interval, then the ∂_{j-1}^{j} should be approximately equal. The nonparametric Friedman test for repeated measurements was used to test for equality of the ∂_{j-1}^{j}. Neither under the linear nor under the power approximation could the null hypothesis of equality of the ∂_{j-1}^{j} be rejected ($\chi^2(3) = 6.270$, $p = 0.099$ and $\chi^2(3) = 2.655$, $p = 0.448$ for the linear and power approximation, respectively). This suggests linearity of the probability weighting function and thus, no systematic deviations from expected utility for intermediate probabilities.

Parametric Weighting Functions. We estimated for each subject the optimal values of the parameters in Equations (2)–(5). The estimation criterion was the minimization of the sum of the squared residuals (SSR), $\sum_{i=1}^{5}(w_i - \hat{w}_i)^2$, where w_i is the elicited probability weight and $\hat{w}_i$ the estimated probability weight under the parametric specification. To be efficient, this estimation criterion requires the error terms to be normally and independently distributed.

Table 4 **Mean Estimation Results for the Parametric Specifications of the Probability Weighting Function**

Functional Form	Linear Approximation	Power Approximation
$w(p) = \frac{p^\gamma}{[p^\gamma + (1-p)^\gamma]^{\frac{1}{\gamma}}}$	$\gamma = 0.713$ (0.025)	$\gamma = 0.674$ (0.027)
$w(p) = \frac{\delta p^\gamma}{\delta p^\gamma + (1-p)^\gamma}$	$\gamma = 0.573$ (0.041)	$\gamma = 0.550$ (0.036)
	$\delta = 1.127$ (0.093)	$\delta = 0.816$ (0.035)
$w(p) = \exp(-(-\ln p)^\alpha)$	$\alpha = 0.589$ (0.037)	$\alpha = 0.533$ (0.031)
$w(p) = \exp(-\beta(-\ln p)^\alpha)$	$\alpha = 0.604$ (0.053)	$\alpha = 0.534$ (0.038)
	$\beta = 0.938$ (0.038)	$\beta = 1.083$ (0.036)

Note. Standard errors appear in parentheses.

Table 4 shows the means of the individual optimal values. For the one-parameter specification Equation (2), our estimates are comparable to those obtained for monetary outcomes. For the two-parameter specification, we find relatively more elevation, in particular under the linear approximation. An explanation for this finding may be that subjects consider all durations in the experiment as losses, because they fall below subjects' life expectancy. It is well-known that people find the chance domain more attractive for losses than for gains (Tversky and Kahneman 1992, Fennema and Wakker 1997). Compared to Tversky and Fox (1995), Wu and Gonzalez (1996), and Abdellaoui (2000), we find more curvature. An explanation for this finding may be that health is a more ''affect-rich'' outcome than money. Rottenstreich and Hsee (1999) find more curvature of the probability weighting function for affect-rich outcomes. However, compared to Gonzalez and Wu (1999), we find less curvature.

Goodness of fit of the various specifications was assessed by taking the mean of the individual sums of squared residuals adjusted for degrees of freedom. Based on this criterion, the two-parameter specifications Equations (3) and (5) fit the data better than their one-parameter counterparts Equations (2) and (4).[2]

[2] Equation (3) leads to reductions in the mean of the individual sums of squared residuals adjusted for degrees of freedom of 34.0% and 24.2% under the linear and the power approximation, respectively, compared to Equation (2). Equation (5) leads to reductions of 26.5% and 11.1% under the linear and the power approximation, respectively, compared to Equation (4).

This better fit of the two-parameter specifications holds in spite of the fact that on the aggregate level both δ and β are relatively close to one and are not significantly different from one under the linear approximation. Apparently, the two-parameter specification fits the data better than the one-parameter specification at the individual subject level, but only slightly better at the aggregate level. This finding is consistent with Gonzalez and Wu (1999).

4.3.3. Propagation of Error. To obtain more insight into the effect of error propagation, we performed two simulation studies based on two different error theories. In the first simulation, we assumed that in evaluating the trade-off questions, the subject makes an error in his assessment of utility differences. This error theory is comparable to Hey and Orme (1994). We assumed that the response error ε is a proportion of the true utility difference, i.e., the assessed utility difference is equal to $(1+\varepsilon)$ times the true utility difference.

In the second simulation, we assumed that while the subject correctly assesses utility differences, he makes an error in reporting his response. This is, by accident the subject sometimes reports the wrong indifference value. This error model is comparable to Harless and Camerer's ''trembling hand'' theory (Harless and Camerer 1994). We assumed that the response error ε is a proportion of the true indifference life duration. That is, the reported indifference life duration is equal to $(1+\varepsilon)$ times the true indifference life duration.

The error terms were in both simulations assumed to be normally distributed, with mean 0.00 and standard deviation 0.05. The selected value of the standard deviation is not important. The aim of the simulation exercise is to show that error propagation is not a problem in our data, that is, that small response errors do not translate into large errors in the elicited probability weights. For both error models, we performed 1,000 simulations. Under both error theories, there is no indication that error propagation is a problem for our elicitation procedure. To illustrate, Table 5 shows the standard deviations of the errors in the aggregate probability weights and, in parentheses, the standard deviations of the errors as a proportion of the probability weights under the linear approximation. In each case, the standard deviation is less than 0.05, the selected size of the response error.

Table 5 **Results of the Simulation Studies in Which There Is an Error in the Assessed Utility Difference (Model I) or in the Reported Indifference Life Duration (Model II)**

Probability	Standard Deviation Error Model I	Standard Deviation Error Model II
0.10	0.0019 (0.8%)	0.0117 (4.8%)
0.25	0.0023 (0.6%)	0.0092 (2.6%)
0.50	0.0024 (0.5%)	0.0057 (1.1%)
0.75	0.0022 (0.3%)	0.0098 (1.5%)
0.90	0.0019 (0.3%)	0.0128 (1.8%)

Notes. Standard deviations of the errors as a proportion of the probability weights are in parentheses. The table shows the error in the aggregate probability weights under the linear approximation.

5. Conclusion

The main conclusion of this article is that probability weighting is robust. We find significant evidence of probability weighting both at the aggregate level and at the individual subject level. The predominant shape of the probability weighting function is inverse S-shaped with the point of inflection lying between 0.25 and 0.50. Probability weighting is particularly strong at the boundaries of the unit interval. Hence, we observe strong support for lower subadditivity and upper subadditivity (Tversky and Wakker 1995). These findings are consistent with studies using monetary outcomes. Compared to these studies, we find more elevation of the probability weighting function.

Wakker and Stiggelbout (1995) have shown how probability weighting can lead to biases in health utility measurement. We have shown that probability weighting affects medical decisions. We urge medical decision analysts to incorporate probability weighting in their analyses. The method outlined in this article can be used to elicit the probability weighting function for individual patients. For societal evaluations where elicitation of the probability weighting function is often not feasible, the parametric specifications Equations (2)–(5) with the parameters elicited in this study can be used.

The robustness of the empirical findings on probability weighting indicates its importance. Ignoring probability weighting in modeling decision under risk and in utility measurement is likely to lead to descriptively invalid theories and distorted elicitations.[3]

[3] We are grateful to Peter Wakker, an associate editor, and two referees for their helpful comments on previous drafts. Han Bleichrodt's research was made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

References

Abdellaoui, M. 2000. Parameter-free elicitation of utilities and probability weighting functions. *Management Sci.* **46**(11).

Beattie, J., G. Loomes. 1997. The impact of incentives upon risky choice experiments. *J. Risk and Uncertainty* **14** 155–168.

Camerer, C. 1992. Recent tests of generalizations of expected utility theory. W. Edwards, ed. *Utility: Theories, Measurement and Applications*. Kluwer Academic Publishers, Boston, MA, 207–251.

——, T.-H. Ho. 1994. Nonlinear weighting of probabilities and violations of the betweenness axiom. *J. Risk and Uncertainty* **8** 167–196.

——, R. M. Hogarth. 1999. The effects of financial incentives in experiments: A review and capital-labor-production framework. *J. Risk and Uncertainty* **19** 7–42.

Cohen, M., J. Y. Jaffray. 1988. Preponderence of the certainty effect over probability distortion in decision making under risk. B. R. Munier, ed. *Risk, Decision and Rationality*. Reidel, Dordrecht, Germany, 173–187.

Currim, I. S., R. K. Sarin. 1989. Prospect versus utility. *Management Sci.* **35** 22–41.

Delquié, P. 1993. Inconsistent trade-offs between attributes: New evidence in preference assessment biases. *Management Sci.* **39** 1382–1395.

——, 1997. ''Bi-matching'': A new preference assessment method to reduce compatibility effects. *Management Sci.* **43** 640–658.

Dubourg, W. R., M. W. Jones-Lee, G. Loomes. 1994. Imprecise preferences and the WTP-WTA disparity. *J. Risk and Uncertainty* **9** 115–133.

Fennema, H. P., P. P. Wakker. 1997. Original and cumulative prospect theory: A discussion of empirical differences. *J. Behavioral Decision Making* **10** 53–64.

Goldstein, W. M., H. J. Einhorn. 1987. Expression theory and the preference reversal phenomena. *Psych. Rev.* **94** 236–254.

Gonzalez, R., G. Wu. 1999. On the form of the probability weighting function. *Cogn. Psych.* **38** 129–166.

Harless, D., C. F. Camerer. 1994. The predictive utility of generalized expected utility theories. *Econometrica* **62** 1251–1289.

Heath, C., A. Tversky. 1991. Preference and belief: Ambiguity and competence in choice under uncertainty. *J. Risk and Uncertainty* **4** 5–28.

Hey, J. D., C. Orme. 1994. Investigating generalizations of expected utility theory using experimental data. *Econometrica* **62** 1291–1326.

Kilka, M., M. Weber. 1998. What determines the shape of the probability weighting function under uncertainty? Working paper, University of Mannheim, Mannheim, Germany.

Lattimore, P. M., J. R. Baker, A. D. Witte. 1992. The influence of probability on risky choice. *J. Econom. Behavior and Organ.* **17** 377–400.

Luce, R. D., P. C. Fishburn. 1991. Rank- and sign-dependent linear utility models for finite first-order gambles. *J. Risk and Uncertainty* **4** 29–59.

McCord, M., R. de Neufville. 1986. Lottery equivalents: Reduction of the certainty effect problem in utility assessment. *Management Sci.* **32** 56–60.

Pliskin, J. S., D. S. Shepard, M. C. Weinstein. 1980. Utility functions for life years and health status. *Oper. Res.* **28** 206–223.

Prelec, D. 1998. The probability weighting function. *Econometrica* **66** 497–528.

Quiggin, J. 1981. Risk perception and risk aversion among Australian farmers. *Australian J. Agricultural Econom.* **25** 160–169.

Rottenstreich, Y., C. K. Hsee. 1999. Money, kisses, and electric shocks: On the affective psychology of risk. Working paper, University of Chicago, Chicago, IL.

Starmer, C., R. Sugden. 1989. Violations of the independence axiom in common ratio problems: An experimental test of some competing hypotheses. *Ann. Oper. Res.* **19** 79–101.

Stiggelbout, A. M., G. M. Kiebert, J. Kievit, J. W. H. Leer, G. Stoter, J. C. J. M. de Haes. 1994. Utility assessment in cancer patients: Adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. *Medical Decision Making* **14** 82–90.

Tversky, A., C. Fox. 1995. Weighing risk and uncertainty. *Psych. Rev.* **102** 269–283.

——, D. Kahneman. 1992. Advances in prospect theory: cumulative representation of uncertainty. *J. Risk and Uncertainty* **5** 297–323.

——, S. Sattath, P. Slovic. 1988. Contingent weighting in judgment and choice. *Psych. Rev.* **95** 371–384.

——, P. P. Wakker. 1995. Risk attitudes and decision weights. *Econometrica* **63** 297–323.

Wakker, P. P., D. Deneffe. 1996. Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. *Management Sci.* **42** 1131–1150.

——, A. M. Stiggelbout. 1995. Explaining distortions in utility elicitation through the rank-dependent model for risky choices. *Medical Decision Making* **15** 180–186.

Wu, G., R. Gonzalez. 1996. Curvature of the probability weighting function. *Management Sci.* **42** 1676–1690.

——, ——. 1998. Common consequence conditions in decision making under risk. *J. Risk and Uncertainty* **16** 115–139.

Yaari, M. E. 1987. The dual theory of choice under risk. *Econometrica* **55** 95–115.

Accepted by Martin Weber. This paper was with the authors 22 weeks for 4 revisions.