Journal of Mathematical Psychology **43**, 238–260 (1999)
Article ID jmps.1999.1257, available online at http://www.idealibrary.com on IDEAL®

Probability Weighting and Utility Curvature in QALY-Based Decision Making

Han Bleichrodt and Jaco van Rijn

iMTA/iBMG, Erasmus University, Rotterdam, The Netherlands

and

Magnus Johannesson

Stockholm School of Economics, Stockholm, Sweden

Quality-Adjusted Life-Years (QALYs) are currently the most important utility model in medical decision making. QALYs are calculated by adjusting years of life for the utility of the health state in which these years are spent. For normative reasons the standard gamble is the preferred method to measure health state utilities, but concern exists about its descriptive properties. Recent theoretical work has suggested that probability weighting can explain anomalies in standard gamble measurement. This paper shows that applying probability weighting in standard gamble measurement increases the consistency of QALYs with individual preferences. The consistency of QALYs with individual preferences is not significantly increased further if utility curvature is also taken into account.

Key Words: QALYs; choice under uncertainty; utility theory.

1. INTRODUCTION

Quality-Adjusted Life-Years (QALYs) are currently the most important outcome measure in health economics and medical decision making. They are important both as a utility-based outcome measure in social decisions about the allocation of health care resources and as a utility model in individual decisions about the selection of the appropriate medical treatment. QALYs offer the advantages of tractability and intuitive appeal, but an important disadvantage is that they are only

We are grateful to Peter Wakker, Elke Weber, and two anonymous referees for helpful comments and suggestions on previous drafts, and to Maiwenn Al for her assistance with the analysis of the results. Han Bleichrodt's research was made possible by a fellowship from the Royal Netherlands Academy of Arts and Sciences. This research was financially supported by the National Corporation of Swedish Pharmacies and by Merck, Sharp and Dohme.

Address correspondence and reprint requests to Han Bleichrodt, iMTA, Erasmus University, P. O. Box 1738, 3000 DR Rotterdam, The Netherlands. E-mail: bleichrodt@bmg.eur.nl.

0022-2496/99 $30.00

under restrictive conditions representative of individual utilities. Characterizations of QALYs have been given for expected utility theory by Pliskin, Shepard, & Weinstein (1980), Bleichrodt (1995), and Bleichrodt, Wakker, & Johannesson (1997), and for a more general rank dependent utility model by Bleichrodt and Quiggin (1997).

The number of QALYs of a health profile (a sequence of health outcomes) is calculated by adjusting each year of life by a quality weight that reflects the attractiveness of the health state in which that year is spent. Formally, if $(q_1, ..., q_T)$ denotes a health profile, where q_t stands for health status in year t, and T is the duration of the profile, then the number of QALYs is equal to

$$\sum_{t=1}^{T} V(q_t), \tag{1}$$

where $V(q_t)$ is the quality weight, or utility, of health state q_t. There exist three principal methods to elicit health state utilities: the rating scale, the time trade-off, and the standard gamble (for a detailed description see Torrance, 1986). The difference between the standard gamble and the other two methods is that the latter do not explicitly involve risk. Risk is a central aspect in most medical decision contexts and it is for this reason that the standard gamble has been advanced as the preferred method of health utility measurement.

In the standard gamble, the analyst asks a subject to compare a risky treatment (t years in health state x, p; t years in health state z), that is, a treatment giving the outcome t years in health state x with probability p and t years in health state z with probability $1-p$, with a certain outcome t years in health state y.[1] The subject is then asked to specify the probability p' for which he is indifferent between the risky treatment and the certain outcome. If we evaluate this indifference by expected utility theory then we obtain

$$\begin{aligned} U(t \text{ years in health state } y) = p'\,U(t \text{ years in health state } x) \\ + (1-p')\,U(t \text{ years in health state } z). \end{aligned} \tag{2}$$

It follows from Eq. (2) after application of the QALY model (Eq. (1)) and the scaling $U(1 \text{ year in health state } x) = 1$ and $U(1 \text{ year in health state } z) = 0$ that $V(\text{health state } y) = p'$.

If a person behaves perfectly well in agreement with expected utility theory then the standard gamble will elicit true utilities. Medical decision making is essentially a prescriptive analysis and expected utility theory is widely seen as the dominant prescriptive model of choice under risk (e.g., Edwards, 1992). However, the elicitation of health state utilities for prescriptive analyses requires subjective judgments and therefore any prescriptive analysis has to take into account descriptive failures

[1] There are other ways of asking standard gamble questions, for example through certainty equivalences or lottery equivalences (Farquhar, 1984; McCord & de Neufville, 1986). However, these methods have not been used to value health states. The probability equivalence method described here is common in health utility measurement and we therefore simply refer to it as the standard gamble method.

of the theory that underlies it. It is now widely acknowledged that expected utility theory fails as a descriptive theory of choice under risk. No direct evidence of descriptive violations of the axioms of expected utility theory exists in the medical decision making literature, but several studies report indirect evidence (e.g., Llewellyn-Thomas *et al*., 1982; Stiggelbout *et al*., 1994). The displayed violations of expected utility theory and anomalies of the standard gamble method have created a deadlock in health utility measurement: on the one hand one would like to stick to the standard gamble because of its normative status, on the other hand the negative empirical evidence has undermined the faith in the validity of the utilities elicited by the standard gamble. This deadlock has led some practitioners in health utility measurement to advocate the use of the time trade-off (Richardson, 1994) and the rating scale (Broome, 1993) instead, even though these methods do not explicitly involve risk and therefore lack validity in decision under risk. From a practical point of view there is clearly a need for procedures to improve the descriptive validity of standard gamble measurement of health state utilities.

In theoretical research on decision under risk, several nonexpected utility theories have been developed in response to the displayed violations of expected utility theory. Among these nonexpected utility theories, the rank dependent utility model (Quiggin, 1982; Yaari, 1987; Schmeidler, 1989) and its derivative cumulative prospect theory are currently the most popular descriptive alternatives for expected utility theory. A distinctive feature of these theories in comparison with expected utility theory is that they allow for nonlinear weighting of probabilities, which empirical evidence has identified as a major cause of violations of expected utility theory.

Contrary to expected utility, in rank dependent utility theory the decision weight of an outcome is not just the probability associated with the outcome, but it is a function of both the probability and the rank of the outcome as compared to the other outcomes. Weber (1994) has argued that rank dependent weighting is applied both as a result of perceptual biases and for motivational reasons. Perceptual biases occur because persons misconceive probabilities. The motivational reasons for rank dependent weighting can either be *individual-specific*, individuals differ in the relative emphasis they put on outcomes at the low end versus outcomes at the high end (see also Lopes, 1984), or *situation-specific*. Situation-specific reasons for rank dependent utility arise because the consequences of over- versus underestimation of risky outcomes may be quite different.[2] Weber and Kirsner (1997) showed that all three reasons for rank dependent utility affect preferences over risky (monetary) outcomes. This is an important finding: if rank dependent weighting would only occur because of perceptual biases then rank dependent utilities would misrepresent true utilities and would be irrelevant in prescriptive analyses. However, if rank dependent weighting is a response to internal and external constraints then it need not be a misrepresentation of true utility and it may have important prescriptive implications.

[2] See also Birnbaum *et al.* (1992), who showed that rank dependent weighting can be derived by minimizing an asymmetric loss function for over- versus underestimates of the value of risky outcomes.

The aim of this paper is to examine whether rank dependent weighting can improve the descriptive validity of the standard gamble in QALY-based decision making. We examine whether the theoretical insights that have been developed in the field of choice under risk can be used to address an important practical problem in health utility measurement: how the descriptive validity of the standard gamble can be improved without sacrificing its normative status. The work by Weber (1994) and Weber & Kirsner (1997) quoted above suggests that rank dependent weighting is also a consequence of motivational factors that are important for prescriptive research. Hence, rank-dependent weighting may indeed be a fruitful way to improve the descriptive validity of standard gamble measurement without sacrificing its normative status.

We compare the descriptive validity of the standard gamble with and without rank-dependent weighting by assessing the degree of association with individual choices. We took individual choices to be the norm, because choices are the basic primitive in decision theory. Any theory of decision under risk takes an individual preference relation as primitive and then proceeds by imposing axioms on this preference relation.

Throughout most of the paper we examine the descriptive validity of standard gambles with and without rank dependent weighting in the context of the QALY model (Eq. (1)). Our motivation to stick to the QALY model is that this model is by far the most important in medical decision making. This paper was motivated by the need to try and develop procedures that help practitioners in health utility measurement to improve the validity of their measurements. Given this aim, it is important to stay as close as possible to the procedures that are commonly applied in health utility measurement. As a consequence, our study provides insight in whether QALYs with rank dependent weighting are more consistent with individual preferences than QALYs without rank dependent weighting. It should be emphasized that our study is not a direct test between rank dependent utility theory and expected utility theory in the medical context. As we observed above, QALYs are only under restrictive assumptions equal to utilities and violation of these assumptions confounds a direct test of rank dependent utility theory versus expected utility theory. We provide some insight in the impact of the confounding by also examining more general QALY type models.

We generalize the QALY model of Eq. (1) by allowing the utility function for life-years to be nonlinear. For *chronic health profiles*, that is, profiles in which health status is constant, we replace the QALY model by the multiplicative representation

$$U(q_1, ..., q_T) = W(t)^* \; V(q). \tag{3}$$

This representation follows both in the expected utility model and in the rank dependent utility model from utility independence, either of quality of life from quantity of life, or of quantity of life from quality of life, and a, in the medical context entirely plausible, zero-condition which says that for a time duration of zero life-years all health states are equivalent (Miyamoto *et al.*, 1998). Empirical analyses have provided support both for utility independence of quantity of life

from quality of life (Miyamoto & Eraker, 1988) and for utility independence of quality of life from quantity of life (Bleichrodt & Johannesson, 1997).

For nonchronic health profiles, profiles in which health status can vary, we replace Eq. (1) by the utility representation

$$U(q_1, ..., q_T) = W(1)^* V(1) + \cdots + W(T)^* V(T). \tag{4}$$

This representation can be characterized by additive independence (Fishburn, 1965) if expected utility theory holds or by generalized marginality (Miyamoto, 1988; Bleichrodt & Quiggin, 1997) if expected utility theory is replaced by a more general rank dependent utility model. Empirical evidence is scarce but negative on additive independence (Maas & Wakker, 1994) and nonexistent on generalized marginality.

In what follows, we briefly describe in Section 2 the difference between expected utility theory and rank dependent utility theory and we show how utilities for health states can be calculated under rank dependent utility theory. In Section 3 we describe the experiment that aimed to test the principal question of this paper: whether rank dependent weighting improves the consistency of QALYs with individual preferences. The results described in Section 4 indicate that this question can be answered in the affirmative: rank dependent weighting indeed leads to a significant improvement in the consistency of QALYs with individual preferences. In Section 5 we move on to the study of the effect of using more general QALY models than Eq. (1) in which curvature of the utility function for life-years is allowed. Section 6 discusses the main implications of our study and concludes.

2. THEORY

We confine ourselves to the context of decision under risk, that is, probabilities are objectively given. Denote a lottery giving outcome x_i with probability p_i by $(p_1, x_1; ...; p_m, x_m)$, where m can be any positive integer. The expected utility of this lottery is equal to

$$\mathrm{EU}(p_1, x_1; ...; p_m, x_m) = \sum_{i=1}^{m} p_i U(x_i), \tag{5}$$

where U is a utility function over outcomes that is not restricted to have any particular shape. Equation (1) shows that in expected utility theory outcomes are transformed into utilities, but probabilities are not transformed and enter linearly in the evaluation formula.

To calculate the rank dependent utility of the lottery $(p_1, x_1; ...; p_m, x_m)$ the outcomes must be rank ordered. Let $x_1 \gtrsim \cdots \gtrsim x_m$, where $\gtrsim$ denotes "at least as preferred as." The rank dependent utility of $(p_1, x_1; ...; p_m, x_m)$ is equal to

$$\mathrm{RDU}(p_1, x_1; ...; p_m, x_m) = \sum_{i=1}^{m} \pi_i U(x_i) \tag{6}$$

and the decision weights π_i are computed as

$$\pi_j = w\left(\sum_{i=1}^{j} p_i\right) - w\left(\sum_{i=1}^{j-1} p_i\right), \tag{7}$$

where w is a probability weighting function (with $w(0)=0$, $w(1)=1$); monotonic (if $a>b$ then $w(a)>w(b)$), but not necessarily additive ($w(a+b)\neq w(a)+w(b)$ can happen). If the probability weighting function is additive ($w(a+b)=w(a)+w(b)$), then rank dependent utility theory reduces to expected utility theory. Equations (6) and (7) show that rank dependent weighting is captured by making the probability weights dependent on the outcome distribution. This distinguishes rank dependent utility from for example lottery dependent utility (Becker & Sarin, 1987), where the outcome weighting is dependent on the probability distribution. To make this distinction clear and to avoid any ambiguities, we use throughout the remainder of the paper the term "probability weighting" instead of "rank dependent weighting."

Based on empirical research, several authors[3] have argued that the probability weighting function w has an inverse S-shaped form, which starts off concave (i.e., it overweights low probabilities) and then becomes convex (i.e., it underweights intermediate and high probabilities). In particular, Tversky & Kahneman (1992) have suggested a parsimonious one parameter functional form:[4]

$$w(p) = \frac{p^{\gamma}}{[p^{\gamma} + (1-p)^{\gamma}]^{1/\gamma}}. \tag{8}$$

If $\gamma = 1$ then $w(p) = p$ and expected utility theory results. For $0.27 < \gamma < 1$, Eq. (4) produces the desired inverse S shape. Tversky & Kahneman (1992) found that γ is equal to 0.61 for gains and to 0.69 for losses. Slightly different estimates of γ were found by Camerer & Ho (1994) ($\gamma = 0.56$) and by Wu & Gonzalez (1996) ($\gamma = 0.71$). In the latter two studies only gains were used in the experiments. These estimates imply that the value of p for which the shape of $w(p)$ changes from concave to convex lies between 0.30 and 0.40.

We explained the standard gamble procedure in the Introduction and we showed there that under expected utility theory, the QALY model, and the scaling U(1 year in health state x) = 1 and U(1 year in health state z) = 0, the utility of health state y is equal to the indifference probability p'. If we evaluate the standard gamble by rank dependent utility (Eq. (6)) instead of expected utility then given the QALY model and the above scaling it follows that $V(y) = w(p')$. The Tversky & Kahneman probability weighting function (Eq. (8)) with γ equal to one of the values found in previous studies implies that if the indifference probability is lower than approximately 0.35, health state utilities with probability weighting will be higher

[3] For example, Quiggin (1982), Karni & Safra (1990), Tversky & Kahneman (1992), Tversky & Wakker (1995).

[4] Prelec (1998) proposed an alternative one parameter functional form for the inverse S shape: $w(p) = \exp(-(-\ln p)^{\gamma})$. As in Tversky and Kahneman's function, if $\gamma = 1$, $w(p) = p$ and if $\gamma < 1$ the function becomes more regressive. Wu & Gonzalez (1996) found that Prelec's function fitted worse than Tversky & Kahneman's.

than health state utilities without probability weighting; if the indifference probability is higher than approximately 0.35, health state utilities with probability weighting will be lower than health state utilities without probability weighting.

3. EXPERIMENT

3.1. Aim

The experiment aimed to compare the descriptive validity of QALYs with probability weighting with QALYs without probability weighting. Descriptive validity is assessed by a comparison with individual preferences which were elicited by a direct ranking task described below.

3.2. Subjects

Eighty student at the Stockholm School of Economics and 92 students at the Erasmus University Rotterdam participated in the study. They were offered $15 for their participation.

3.3. Procedure

The experiment was carried out in different sessions with on average 10 respondents per session. Before the experiment was carried out, the questionnaire was tested both in Stockholm and in Rotterdam with faculty staff members as respondents.

3.4. Health State

The health state we selected corresponds to a severe type of lower back pain. We chose a fairly common health problem to make it easier for respondents to imagine the problem. We described the health state by level of functioning on four attributes: general daily activities, self care, leisure activities, and pain. Table 1

TABLE 1

The Health States "Severe Lower Back Pain" and "Full Health"

Severe lower back pain
• Unable to perform some tasks at home and/or at work
• Able to perform all self cares activities (eating, washing, dressing) albeit with some difficulties
• Unable to participate in many types of leisure activity
• Often moderate to severe pain and/or other complaints
Full health
• Able to perform all tasks at home and/or at work without problems
• Able to perform all self care activities (eating, washing, dressing) without problems
• Able to participate in all types of leisure activity without problems
• No pain or other complaints

describes the selected health state to which we refer in the remainder of this paper as "severe lower back pain." Table 1 also describes the health state "full health," which was defined as no limitations on each of the four attributes.

3.5. *Task*

Subjects were faced with a standard gamble question in which the certain option was living for 30 years with severe lower back pain and the treatment option was (30 years in full health, p; immediate death). They were told that all profiles were followed by death after 30 years. Subjects had to indicate the value of p for which they were indifferent between these two options on a line of values for p, calibrated between 0 and 1 with ticks 0.01 point apart. Next to this line a line was drawn with the corresponding value of $1-p$. This was done to remind respondents what a choice of p implied in terms of the probability of immediate death. By this procedure, we hoped to control for a potential framing bias: only displaying the probability of successful treatment might lead to too strong a focus on the outcome of successful treatment, full health.

Subjects were encouraged to use a bounding method in answering the questionnaire: they were asked first to indicate those values of p for which they definitely preferred 30 years with severe lower back pain, then those values of p for which they definitely preferred the treatment option, and finally those values of p for which they did not have a preference for one of the options. We encouraged subjects to make this latter range of values as small as possible, but we did not force them to state just one indifference value. Subjects are not used to express gains in quality of life on a probability scale and most of them had never actually experienced severe lower back pain. Determining preferences by an unfamiliar method, in a limited time period, and for a health state subjects have not lived through, may cause preferences to be somewhat imprecise. We thought it better to allow subjects to express this imprecision of preference. In the analyses we used the midpoint of the given range of values. To examine the sensitivity of the results we performed separate analyzes with the upper limit and the lower limit of the range of values.

After they had completed the standard gamble question, subjects were asked to rank seven health profiles. The health profiles were printed on separate cards and were presented in random order. Table 2 describes the seven profiles. All profiles were followed by death. If profiles consisted of both years with severe lower back pain and years in full health, then the years in full health always came first. We learned from pilot sessions that profiles of decreasing quality of life were more in line with people's expectations and therefore easier to process. The rank order of several of the profiles is self-evident. If subjects' preferences are monotonic with respect to life-years, in the direction that more life years are preferred, then the rank order of profiles 2–5 is obvious. We included all these profiles in the experiment to ensure enough variation in the ranking of the profiles over a wide range of quality weights.

The aim of the ranking exercise was to elicit individual preferences directly. The ranking section was always performed after the standard gamble assessment. We chose this order of the tasks because we thought that it might increase the reliability of the data. To compare the profiles in the ranking task, subjects must make trade-offs between quality of life and quantity of life. These trade-offs are

TABLE 2

The Seven Health Profiles Used in the Experiment

Number profile	Years in full health	Years with severe lower back pain
1	0	20
2	18	0
3	16	0
4	14	0
5	12	0
6	8	8
7	6	11

easier to make if subjects have an impression of the difference in utility between full health and severe lower back pain. By doing the standard gamble assessment, where the time dimension is held fixed, first, we hoped that subjects would get a better understanding of the utility difference between severe lower back pain and full health.

3.6. Methods

From the indifference value p elicited by the standard gamble question we can calculate the utility of severe back pain both with and without probability weighting. We used the Tversky & Kahneman probability weighting function (Eq. (8)) to compute the utility of severe back pain with probability weighting. The utility of severe back pain was then used to compute for each of the profiles the number of QALYs with and without probability weighting. We calculated for each profile and for each subject the number of QALYs with probability weighting for all theoretically allowed values of γ ($0.27 \leqslant \gamma \leqslant 1$). For each subject and for each value of γ the predicted QALY ranking was then compared with the ranking that was elicited directly. We assessed the strength of the association between the predicted QALY ranking and the direct ranking by the Spearman rank correlation coefficient. Finally, we determined the value of γ for which the Spearman rank correlation coefficient was maximal and we compared this maximal value with the value of the Spearman rank correlation coefficient when no probability weighting was applied. We determined the maximizing value of γ by two approaches. In the first approach we excluded individual-specific differences and we used just one value of γ for all subjects. That is, for each value of γ we determined for each subject the Spearman rank correlation coefficient and we then averaged these over all subjects. This procedure is most relevant to the use of QALYs in societal decisions on resource allocation where one is interested in the preferences of the "representative individual" and individual-specific differences are of less importance. In the second approach we determined for each individual separately the value of γ for which the Spearman rank correlation coefficient was maximized and we then calculated the average of these maximized Spearman rank correlation coefficients. This procedure is most relevant for individual medical decision making where individual differences are clearly important.

4. RESULTS

We present the results for the total sample. The results did not differ significantly between the Dutch and the Swedish samples. It further turned out that the results are not sensitive to whether the lower limits, the middle points or the upper limits of the ranges of indifference probabilities are used in the analysis. We therefore only present the results based on the middle points of the ranges of indifference probabilities.

Figure 1 shows the distribution of the indifference probabilities elicited by the standard gamble procedure. The indifference probabilities range between 0 and 1, with a mean of 0.668 (standard error = 0.019) and a median of 0.715. Given the estimates of the probability weighting parameter γ found by Camerer & Ho (1994), Tversky & Kahneman (1992), and Wu & Gonzalez (1996), most respondents are on the convex part of the probability weighting function.

Figure 2 gives the results for the procedure in which individual-specific differences were ignored. The figure shows the graph of the average Spearman rank correlation coefficient as a function of γ. The graph reaches its maximum at $\gamma = 0.69$ (RCC = 0.809). The figure also shows that the rank correlation coefficients do not vary much over a rather wide range of values of γ. In fact, the rank correlation coefficients do not differ significantly ($\alpha = 0.05$) for values of γ between 0.54 and 0.72. This range includes the estimates of Camerer & Ho (1994) (0.56), Tversky & Kahneman (1992) (0.61 for gains and 0.69 for losses), and Wu & Gonzalez (1996) (0.71).

Even if we ignore individual-specific differences, QALYs with probability weighting ($\gamma = 0.69$) are significantly more consistent with the directly elicited ranking than QALYs without probability weighting: the average rank correlation coefficient rises from 0.730 to 0.809 ($p < 0.001$).

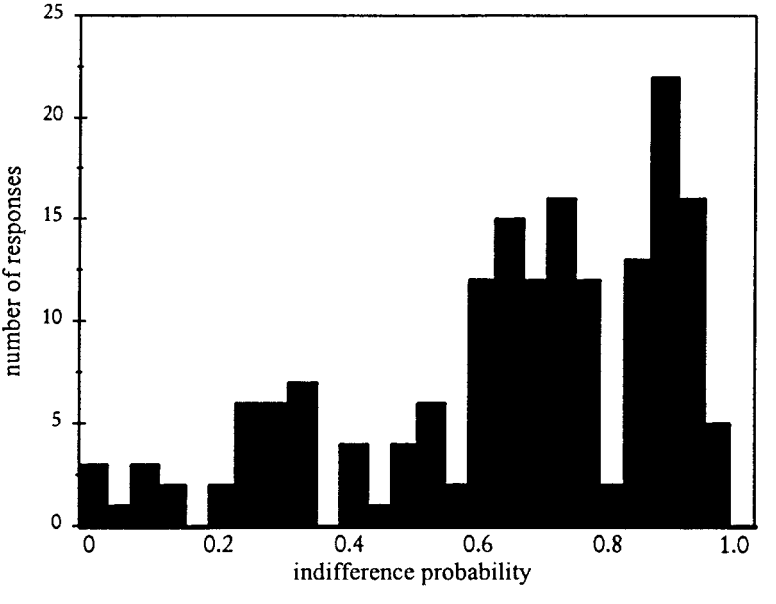

FIG. 1. The distribution of the elicited indifference probabilities.

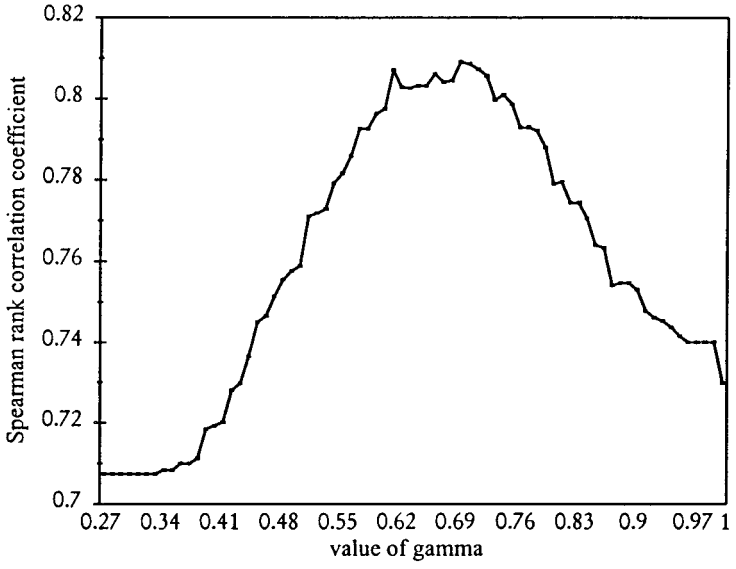

FIG. 2. Mean Spearman rank correlation coefficients as a function of the value of γ.

The above analysis ignores individual-specific differences. However, Gonzalez (1993) has shown that there is substantial across-subject heterogeneity in weighting function parameters. We therefore also determined the maximum value when individual-specific differences are taken into account. These data should be interpreted with caution. If monotonicity with respect to years in full health holds, and none of our subjects violated this, then the number of degrees of freedom in each individual estimation is low, which may lead to instable estimations. For most subjects there was a range of values of γ for which the RCC was maximal. The analysis has been based on the midpoint of this range.

Figure 3 displays the distribution of the maximizing value of γ. The figure shows that the distribution is centered around $\gamma = 0.65$. Taking into account individual-specific differences increases the average rank correlation coefficient to 0.937 ($p < 0.001$). The lack of degrees of freedom in the individual estimations led to rather wide ranges of values of γ for which the Spearman rank correlation coefficient was maximized. The average size of the range of maximizing values was 0.30. For 68 subjects $\gamma = 1$ was included in the range of optimal values, which implies that for these subjects probability weighting did not increase the consistency of QALYs with the individual choices.

What is the reason that QALYs with probability weighting are more consistent with the directly elicited ranking? For the maximizing values of γ found above, most subjects are on the convex part of the probability weighting function and therefore underweight probabilities. By consequence, for most subjects probability weighting leads to a lower utility for severe lower back pain. The higher consistency of QALYs with probability weighting indicates that if probability weighting is not taken into account then quality weights will be too high (given common scaling). The standard gamble we used is a probability equivalence method and our results are in line with other studies that found similar overestimation of utilities by probability equivalence methods (Hershey & Schoemaker, 1985; Wakker & Deneffe,

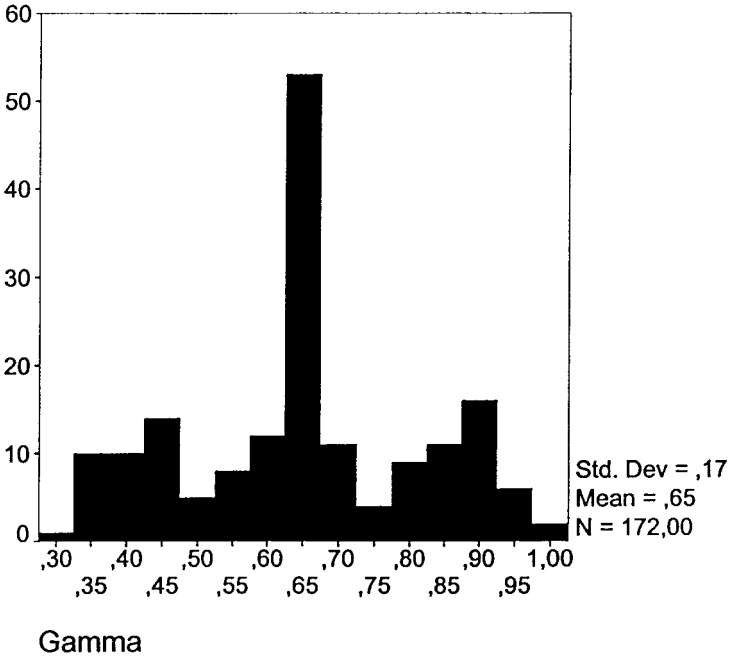

FIG. 3. The distribution of the values of γ for which the Spearman rank correlation coefficient is maximal.

1996). Hence, we are inclined to conclude that the major improvement of probability weighting is that it adjusts an upward bias that is present in probability equivalence methods. This is confirmed if we split the sample in a group that is on the concave part of the weighting function ($p < 0.38$) and in a group that is on the convex part of the weighting function ($p \geqslant 0.38$). In the group that is on the concave part of the weighting function, for which probability weighting leads to a higher utility, the rank correlation coefficients for QALYs with probability weighting and for QALYs without probability weighting are equal[5] ; in the group that is on the convex part of the weighting function, the average rank correlation coefficient for QALYs with probability weighting is about 0.10 higher than the average rank correlation coefficient for QALYs without probability weighting if we ignore individual-specific differences and 0.25 higher if individual-specific differences are taken into account ($p < 0.001$ in both comparisons).

5. UTILITY CURVATURE

As we noted in the Introduction, the results presented in Section 4 may to some extent have been confounded by violations of the conditions underlying the QALY model. We therefore reanalyzed the data replacing the linear utility function for life-years of Eq. (1) by a nonlinear function.

We examined two specifications of the utility function for life-years. The first is the log/power family, $W(t) = t^r$, which has been used in medical decision making by

[5] This holds regardless of whether individual-specific differences are taken into account.

Pliskin *et al.* (1980), Miyamoto & Eraker (1985), and Stiggelbout *et al.* (1994) and for monetary outcomes by Tversky and Kahneman (1992), Camerer & Ho (1994), and Wu & Gonzalez (1996). The second specification is the linear/exponential family $W(t) = e^{-ct}$, which has been frequently applied in decision analyses because it corresponds to constant rate discounting (e.g., Viscusi & Moore, 1989; Moore & Viscusi, 1990).

Pratt (1964) and Miyamoto (1988) have derived simple conditions which characterize these two families of utility functions for expected utility and general rank dependent utility, respectively. Miyamoto & Eraker (1989) tested these conditions and found that even though neither model gave a good description for individual subject data, the linear/exponential utility model was an excellent approximation for data averaged across individuals.

We analyzed the data both with and without allowance for individual-specific differences. We calculated the maximum Spearman rank correlation coefficient both for the model in which the utility function for life-years can be nonlinear, but there is no probability weighting ($QALY_{UC}$) and for the model in which the utility function for life-years can be nonlinear and there is probability weighting ($QALY_{UC+PW}$). In the latter model we had to determine the maximizing values of two coefficients: the coefficient that reflects curvature of the utility function (either r or c) and the coefficient of the probability weighting function (γ).

Tables 3 and 4 show the maximizing rank correlation coefficients when individual-specific differences are ignored. Table 3 displays the results for the log/power family. The range of power coefficients we examined includes the overall estimates from the studies by Stiggelbout *et al.* ($r = 0.74$), by Miyamoto & Eraker (r varies between 0.91 and 1.10 for different age groups), and by Pliskin *et al.*

TABLE 3

Mean Spearman Rank Correlation Coefficients (RCC) for QALYs with Only Utility Curvature but No Probability Weighting ($QALY_{UC}$) and for QALYs with Both Utility Curvature and Probability Weighting ($QALY_{UC+PW}$)

Power coefficient (r)	RCC $QALY_{UC}$	RCC $QALY_{UC+PW}$	Maximizing value of γ
0.25	0.774[a]	0.782[b]	0.93
0.5	0.770[a]	0.793	0.78
0.75	0.767[a]	0.802	0.72
0.9	0.749[a]	0.806	0.64
1.1	0.706[a]	0.813	0.64
1.25	0.678[a]	0.815	0.59
1.5	0.648[a]	0.806	0.53
1.75	0.616[a]	0.785[b]	0.49

Note. The utility function is from the log/power family.

[a] significantly different from RCC of QALYs with only probability weighting ($\gamma = 0.69$) at the 1% significance level.

[b] significantly different from RCC of QALYs with only probability weighting ($\gamma = 0.69$) at the 5% significance level.

TABLE 4

Mean Spearman Rank Correlation Coefficients (RCC) for QALYs with Only Utility Curvature but No Probability Weighting ($QALY_{UC}$) and for QALYs with Both Utility Curvature and Probability Weighting ($QALY_{UC+PW}$)

Exponent (c)	RCC $QALY_{UC}$	RCC $QALY_{UC+PW}$	Maximizing
−0.10	0.581[a]	0.774[a]	0.49
−0.05	0.635[a]	0.812	0.52
−0.03	0.665[a]	0.814	0.58
−0.01	0.706[a]	0.812	0.65
0.01	0.744[a]	0.808	0.63
0.03	0.767[a]	0.803	0.70
0.05	0.776[a]	0.802	0.77
0.10	0.782[a]	0.791	0.83

Note. The utility function is from the linear/exponential family.

[a] significantly different from RCC of QALYs with only probability weighting ($\gamma = 0.69$) at the 1% significance level.

[b] significantly different from RCC of QALYs with only probability weighting ($\gamma = 0.69$) at the 5% significance level.

($r = 1.14$).[6] Table 3 shows that if the utility function for life years is concave ($r < 1$), then QALYs with only utility curvature but no probability weighting are more consistent with the directly elicited ranking than QALYs without both utility curvature and probability weighting. However, for all values of the power coefficient, QALYs with only utility curvature but no probability weighting are less consistent with the directly elicited ranking than QALYs with only probability weighting and no utility curvature ($\gamma = 0.69$). The difference is significant at the 1% level. Further, QALYs with both utility curvature and probability weighting are not significantly more consistent with the directly elicited ranking than QALYs with only probability weighting but no utility curvature ($p > 0.05$). This suggests that the adjustment for probability weighting is more important to improve the consistency of QALY based analyses than the adjustment for utility curvature.

Table 4 shows the results for the linear/exponential family of utility functions. A positive value of the exponent c corresponds to a concave utility function for life-years. Table 4 confirms the above pattern: QALYs with a concave utility function for life-years are more consistent with the direct ranking than QALYs with a linear utility function for life-years, QALYs with only probability weighting are more consistent with the direct ranking than QALYs with only utility curvature, and QALYs with both utility curvature and probability weighting are not significantly more consistent with the direct ranking than QALYs with only probability weighting.

Figures 4 and 5 show the results of the analysis when individual-specific differences are taken into account. The figures show the distributions of the maximizing

[6] Values of r greater than one are counterintuitive: they imply that the utility difference between for example year 20 and year 19 is greater than the utility difference between year 1 and year 0. Finding values of r greater than one may be a consequence of violations of the underlying utility model.

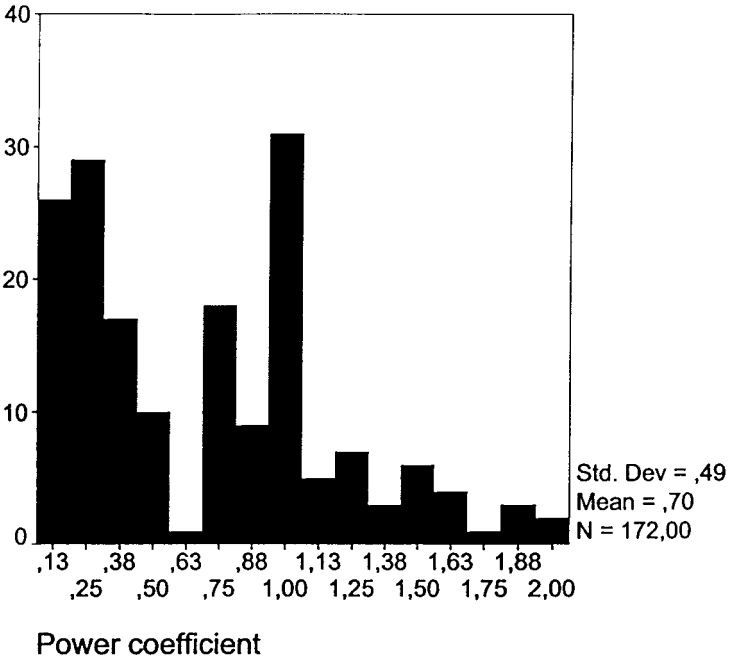

FIG. 4. The distribution of the values of the power r for which the Spearman rank correlation coefficient is maximal.

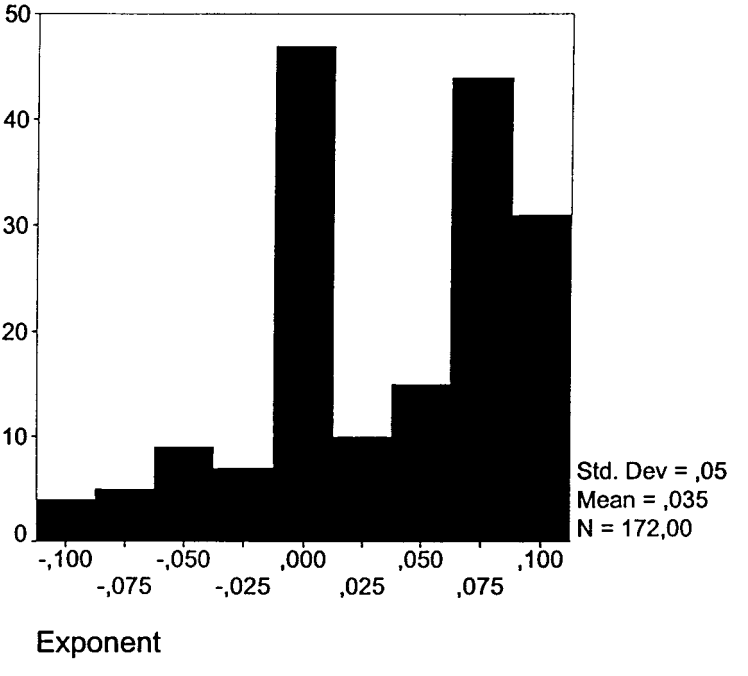

FIG. 5. The distribution of the values of the exponent c for which the Spearman rank correlation coefficient is maximal.

values of the power coefficient and the exponential coefficient respectively for QALYs with utility curvature but without probability weighting. The distributions are essentially two peaked: there is a group of subjects for whom the linear utility function fits best and there is a group of subjects for whom the utility function for life-years is concave. The individual-specific analysis further reveals that the power family is more consistent with the direct ranking than the exponential family: the mean rank correlation coefficient for the power family is equal to 0.900 as opposed to 0.861 for the exponential family. The difference is significant ($p < 0.001$). The individual analysis confirms that QALYs with only probability weighting are more consistent with the direct ranking than QALYs with only utility curvature. The increase in the rank correlation coefficient is larger if we incorporate only probability weighting in the QALY model (from 0.730 to 0.937) than if we incorporate only utility curvature in the QALY model (maximal increase from 0.730 to 0.900). The difference is significant ($p < 0.001$). A further increase in the rank correlation coefficient is obtained by incorporating both probability weighting and utility curvature (power function) in the QALY model: from 0.937 to 0.961. However, the estimates of the coefficients γ and r are very instable: for some subjects there were over 1000 pairs of values for γ and r that maximized the Spearman rank correlation coefficient. This is due to a lack of degrees of freedom in the individual estimations. Given monotonicity with respect to life-years only three profiles are free to vary. In the QALY model with both probability weighting and utility curvature two parameters have to be estimated which leaves only one degree of freedom.

The model $U(q_1, ..., q_T) = V(q_1)^* W(1) + \cdots + V(q_T)^* W(T)$, which we used to analyze nonchronic health states, only represents individual preferences if either additive independence or generalized marginality holds. There exists no empirical support for these conditions and we therefore reanalyzed the data using the multi-

TABLE 5

Mean Spearman Rank Correlation Coefficients (RCC) for QALYs with Only Utility Curvature but No Probability Weighting ($QALY_{UC}$) and for QALYs with Both Utility Curvature and Probability Weighting ($QALY_{UC+PW}$)

Power coefficient (r)	SRCC $QALY_{UC}$	SRCC $QALY_{UC+PW}$	Maximizing value of γ
0.25	0.849[a]	0.855[a]	0.94
0.5	0.840[a]	0.860[a]	0.81
0.75	0.841[a]	0.875	0.71
0.9	0.829[a]	0.883	0.64
0	0.813[a]	0.883	0.61
1.1	0.797[a]	0.884	0.63
1.25	0.781[a]	0.885	0.56
1.5	0.749[a]	0.893	0.53
1.75	0.713[a]	0.893	0.48

Note. The utility function is from the log/power family. Only chronic health profiles are included.

[a] significantly different from RCC of QALYs with only probability weighting ($\gamma = 0.69$) at the 1% significance level.

[b] significantly different from RCC of QALYs with only probability weighting ($\gamma = 0.69$) at the 5% significance level.

plicative model $U(Q, T) = W(T)^* V(Q)$ for which more empirical support exists. Because this model can only be applied to chronic health profiles, we had to exclude profiles 6 and 7 from the analysis. Dropping profiles 6 and 7 means that if monotonicity of preferences with respect to years in full health holds, then the various models only have to explain where profile 1 fits in the rank ordering. So there is only one degree of freedom left in each individual estimation and we therefore only report the results of the analysis in which individual-specific differences are ignored. Tables 5 and 6 display the results for the power and exponential utility functions respectively. Inspection of these tables confirms the pattern observed above: QALYs with only probability weighting are more consistent with the direct ranking than QALYs with only utility curvature; taking both utility curvature and probability weighting into account does not lead to a significant further increase in the consistency of QALYs with the direct ranking.

How can we explain that utility curvature in addition to probability weighting does not further improve the consistency of QALY based decision making? We noted above that the major improvement that is achieved through probability weighting is that less weight is given to years with severe lower back pain. In the context of our experiment utility curvature has the same impact if the utility function for life-years is concave. The profiles we used are all of a decreasing quality: years with severe lower back pain always came after years in full health. Both a utility function with a power coefficient (r) less than 1 and a utility function with a positive exponent (c) give less weight to more distant years, that is, to years with severe lower back pain. In summary, probability weighting and utility curvature have the same impact in the context of our experiment: to decrease the weight given to years with severe lower back pain. However, the results suggest that probability weighting does so in a way that is more consistent with individual preferences.

TABLE 6

Mean Spearman Rank Correlation Coefficients (RCC) for QALYs with Only Utility Curvature but No Probability Weighting ($QALY_{UC}$) and for QALYs with Both Utility Curvature and Probability Weighting ($QALY_{UC+PW}$)

Exponent (c)	RCC $QALY_{UC}$	RCC $QALY_{UC+PW}$	Maximizing value of γ
−0.10	0.673[a]	0.887	0.49
−0.05	0.733[a]	0.892	0.52
−0.03	0.764[a]	0.886	0.56
−0.01	0.797[a]	0.884	0.62
0.01	0.829[a]	0.881	0.69
0.03	0.840[a]	0.877	0.71
0.05	0.851[b]	0.874	0.77
0.10	0.858[a]	0.863[a]	0.86

Note. The utility function is from the linear/exponential family. Only chronic health profiles are included.

[a] significantly different from RCC of QALYs with only probability weighting ($\gamma = 0.69$) at the 1% significance level.

[b] significantly different from RCC of QALYs with only probability weighting ($\gamma = 0.69$) at the 5% significance level.

6. CONCLUSION

The main conclusion of this paper is that incorporating probability weighting in the standard gamble method improves the consistency of QALY-based decision making with individual preferences. Our research was motivated by an existing impasse in health utility measurement: it is widely recognized that the standard gamble is normatively more valid than the other methods of health utility measurement, but the descriptive validity of the standard gamble is troublesome. The observed anomalies in standard gamble measurement have undermined the faith in the method and have contributed to the use of alternative methods that lack theoretical validity in comparison with the standard gamble but that are easier to apply. The lack of descriptive validity of the standard gamble has led researchers in the field of health utility measurement to the conclusion that the standard gamble may be ideal in theory but not very suitable in the practice of health utility measurement. We have shown that a recent insight from the theory on decision under risk, probability weighting, can successfully be applied to increase the consistency of standard gamble utilities with individual preferences. Previous research (Weber, 1994; Weber & Kirsner, 1997) has indicated that probability weighting occurs for motivational reasons. Therefore, probability weighting may be relevant for prescriptive purposes and can be used in health utility measurement to increase the descriptive validity of the standard gamble without sacrificing its normative validity.

QALYs are used both as a utility-based outcome measure in societal decisions about the allocation of health care resources and as a utility model in individual medical decisions concerning the selection of appropriate treatment. Our results are important for both applications of QALYs. If QALYs are used in social decision making, individual-specific differences are less important and interest is primarily focused on the results for the "representative subject." We have shown that a significant improvement in descriptive validity can be obtained if probability weighting is incorporated in QALY-based decision making even when individual-specific differences are ignored. We recommend that in economic evaluations of health care programs where the standard gamble has been used, utilities are adjusted for probability weighting by the Tversky & Kahneman (1992) weighting function (Eq. (8)) with the value of γ equal to 0.69. Individual-specific differences are obviously relevant in medical decisions about the selection of appropriate treatment. Our results indicate that using individual-specific rather than average estimates for the probability weighting parameter γ can further increase the consistency of utility estimates. Individual estimation of the probability weighting parameter requires additional elicitations, but this is worth the effort because it leads to more representative utilities.

Thus far, studies have attempted to improve the consistency of QALYs with individual preferences by focusing on the utility function for life-years. People have used power utility functions and exponential utility functions instead of the linear function of the QALY model (Eq. (1)). Our results imply that a larger gain in consistency can be obtained by probability weighting. This is not to say that one could just stick to the linear utility function. There are some indications in our results that a concave utility function may lead to a further increase in consistency. However,

the results clearly show that only focusing on the utility part of QALYs while ignoring the impact of probability weighting is not a fruitful strategy. Future research on health utility measurement, both theoretical and empirical, should also be directed at the modeling of probability weighting.

One might criticize our study for being overly preoccupied with statistical significance, ignoring the question whether the differences between the average rank correlation coefficients are actually meaningful. It is true that all rank correlation coefficients are high and according to classification schemes (Landis & Koch, 1977) generally fall in the same category. However, this is an artifact of the inclusion of profiles for which the ordering follows naturally from monotonicity of preferences with respect to years in full health. We had to include these profiles to ensure enough variation in the ranking of the profiles over a wide range of quality weights, but a disadvantage of their inclusion is that they inflate the rank correlation coefficients. If we only include profiles for which the ordering does not follow naturally from monotonicity with respect to years in full health, the rank correlation coefficients drop considerably. If we only include profiles 1, 4, 6, and 7, the maximum rank correlation coefficient for QALYs with probability weighting ($\gamma = 0.69$) becomes 0.446 and the rank correlation coefficient for QALYs without probability weighting becomes 0.286; if we only include profiles 1, 5, 6, and 7, the maximum rank correlation coefficient for QALYs with probability weighting ($\gamma = 0.69$) becomes 0.487 and the rank correlation coefficient for QALYs without probability weighting becomes 0.362. The differences between the rank correlation coefficients are significant ($p < 0.01$) and the correlation coefficients fall in different categories of the Landis & Koch classification scheme, which suggests that they are also meaningful.

There are several limitations of our study that may be addressed in future research. First, we implicitly assumed that individual preferences can be measured by simultaneously ranking the seven profiles. A different procedure is to ask subjects to make pairwise choices between the profiles and to derive the ranking of the profiles from these answers. It is not clear whether the two procedures lead to identical rankings and which procedure is to be preferred. On the one hand, pairwise choices may be slightly more in line with the basic primitive of decision theory, preference over pairs of alternatives, on the other hand, a pairwise choice procedure is vulnerable to intransitivities (Tversky, 1969) and intransitivity is highly undesirable from a normative point of view.

Second, few of our subjects had actually ever experienced severe back pain, which utility they had to elicit. Hence, our results are based on predicted utility, which as recent research by Kahneman and others has shown may be different from experienced utility (Fredrickson & Kahneman, 1993; Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993; Varey & Kahneman, 1992; Kahneman, Wakker, & Sarin, 1997). The distinction predicted utility versus experienced utility touches on the discussion whose preferences should count in medical decision making. For individual medical decision making, it is obvious that the preferences of the patient should count, and in that context experienced utility may be more relevant than predicted utility. Societal decisions about which programs to fund affect both patients and nonpatients and in this context the preferences of those who have

never experienced the health states (and thus predicted utility) are relevant. Therefore, our results may be more relevant for societal decision making than for individual decision making and it would be interesting to repeat our experiment in the context of individual medical decision making with patients instead of students as experimental subjects.

The selection of the type of profiles may also have affected our results. We selected profiles of decreasing quality, because we had learned from pilot sessions that subjects found such profiles easier to imagine. However, the selection of profiles of decreasing quality makes that our conclusion about the impact of utility curvature, that concave utility is most consistent with individual preferences, is somewhat tentative. As we explained in Sections 4 and 5, we believe that the reason why probability weighting and concave utility improve the consistency of QALYs with individual preferences is because they decrease the weight given to years with severe back pain. Previous studies have also shown that the use of a probability equivalence method leads to utilities that are too concave (given common scaling). If we use profiles of decreasing quality then concave utility will imply that less weight is given to the utility of severe back pain. However, if profiles of increasing quality are used then concave utility will give extra weight to the years with severe back pain. Therefore, if our hypothesis is true then we should find that for profiles of increasing quality convex utility is more consistent with individual preferences than concave utility.

We do not believe that our conclusions about the impact of probability weighting are sensitive to the selection of the health profiles. However, they may be sensitive to the selection of the health state and it is certainly interesting to examine the robustness of our conclusions in designs that involve other health states. In particular, it is interesting to examine whether our conclusions still hold for health states which are only slightly preferred to death. For such health states we expect the median indifference probability to be smaller than 0.3 and the majority of subjects will be on the concave part of the weighting function. This implies that probability weighting leads to utilities that are higher than the utilities elicited without probability weighting. If the hypothesis that QALYs with probability weighting are more consistent with individual preferences than QALYs without probability weighting because probability weighting adjusts the utility of severe back pain downward is true, then it cannot be excluded that probability weighting will be less consistent with individual preferences for health states that are only slightly more attractive than death.

A final issue that could be addressed in future research is the impact on the results of the mode of assessment of the indifference probabilities. We asked subjects to determine an indifference probability by crossing the probabilities for which they had a clear preference for one of the options, which is a type of matching task. An alternative method is to establish indifference from a series of binary choices. The results from these two procedures need not be equivalent. Bostic, Herrnstein, & Luce (1990) have examined the two procedures and they found that binary choice tasks lead to less inconsistencies in preferences than matching tasks. These findings suggest that our approach of relying on a matching task may have introduced biases in the elicited probabilities that could be avoided by using a choice

task. The work by Tversky, Sattah, & Slovic (1988) on scale compatibility effects shows that the characteristics of the task and of the response scale prime the most compatible features of the stimulus. In the matching task, the response scale is the probability of successful treatment. In answering the gamble question, scale compatibility predicts that people anchor on the probability in the certain outcome, which is equal to one, and then adjust downward. This adjustment is generally insufficient and hence the matching task we used may have led to an overestimation of true utilities.

This paper has shown that there are ways to improve the descriptive validity of standard gamble measurement without sacrificing its normative appeal. This is an important conclusion. Previous research has displayed anomalies of the standard gamble. Two courses of action are possible in response to these anomalies: first, one could seek and develop alternative methods that avoid the inconsistencies in standard gamble measurement; second, one could try to improve the descriptive validity of the standard gamble. The first reaction has been typical in health utility measurement. A disadvantage of this approach is that the theoretical basis of these alternative methods is often unclear and that they lack validity as utilities in decision under risk. Most medical decision making takes place under risk and it is obviously desirable to use utilities that are valid in this context. The standard gamble elicits utilities that are valid in decision under risk and we therefore believe that the second reaction, to improve the descriptive validity of standard gamble measurement, is the preferred course of action. In this paper, we have shown that an important improvement in the descriptive validity of the standard gamble can be obtained by incorporating probability weighting in medical decision making. Incorporating probability weighting in practical applications of QALYs is straightforward. We encourage future research, both theoretical and practical, to find further ways to improve the descriptive validity of standard gamble measurement. It is only through such work that methods can be developed that truly represent individual preferences for health.

REFERENCES

Becker, J. L., & Sarin, R. K. (1987). Lottery dependent utility. *Management Science*, **33**, 1367–1382.

Birnbaum, M., Coffey, G., Mellers, B. A., & Weiss, R. (1992). Utility measurement: Configural weight theory and the judge's point of view. *Journal of Experimental Psychology*: *Human Perception and Performance*, **18**, 331–346.

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

Bleichrodt, H., & Johannesson, M. (1997). The validity of QALYs: An experimental test of constant proportional tradeoff and utility independence. *Medical Decision Making*, **17**, 21–32.

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

Bleichrodt, H., Wakker, P. P., & Johannesson, M. (1997). Characterizing QALYs by risk neutrality. *Journal of Risk and Uncertainty*, **15**, 107–114.

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

Broome, J. (1993). QALYs. *Journal of Public Economics*, **50**, 149–167.

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

Edwards, W. (1992). *Utility theories*: *Measurements and applications*. Boston, MA: Kluwer Academic.

Farquhar, P. (1984). Utility assessment methods. *Management Science*, **30**, 1283–1300.

Fishburn, P. C. (1965). Independence in utility theory with whole product sets. *Operations Research*, **13**, 28–45.

Frederickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology*, **65**, 45–55.

Gonzalez, R. (1993). *Estimating the weighting function*. Paper presented at the 26th Annual Mathematical Psychology Meeting, 1993.

Hershey, J. C., & Schoemaker, P. H. J. (1985). Probability versus certainty equivalence methods in utility measurement: Are they equivalent?. *Management Science*, **13**, 1213–1231.

Kahneman, D., Frederickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science*, **4**, 401–405.

Kahneman, D., Wakker, P. P., & Sarin, R. K. (1997). Back to Bentham? Explorations of experienced utility. *Quarterly Journal of Economics*, **112**, 375–405.

Karni, E., & Safra, Z. (1990). Rank-dependent probabilities. *Economic Journal*, **100**, 487–495.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

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.

Lopes, L. (1984). Risk and distributional inequality. *Journal of Experimental Psychology*: *Human Perception and Performance*, **10**, 465–485.

Maas, A., & Wakker, P. P. (1994). Additive conjoint measurement for multiattribute utility. *Journal of Mathematical Psychology*, **38**, 86–101.

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

Miyamoto, J. (1988). Generic utility theory: Measurement foundations and applications in multiattribute utility theory. *Journal of Mathematical Psychology*, **32**, 357–404.

Miyamoto, J., & Eraker, S. (1985). Parameter estimates for a QALY utility model. *Medical Decision Making*, **5**, 191–213.

Miyamoto, J., & Eraker, S. A. (1988). A multiplicative model of the utility of survival duration and health quality. *Journal of Experimental Psychology*: *General*, **117**, 3–20.

Miyamoto, J., & Eraker, S. A. (1989). Parametric models of the utility of survival duration: Tests of axioms in a generic utility framework. *Organizational Behavior and Human Decision Processes*, **44**, 166–202.

Miyamoto, J., Wakker, P. P., Bleichrodt, H., & Peters, H. (1998). The zero-condition: A simplifying assumption in QALY measurement. *Management Science*, **44**, 839–849.

Moore, M. J., & Viscusi, W. K. (1990). Models for estimating discount rates for long-term health risks using labor market data. *Journal of Risk and Uncertainty*, **3**, 381–401.

Pliskin, J. S., Shepard, D. S., & Weinstein, M. C. (1980). Utility functions for life years and health status. *Operations Research*, **28**, 206–224.

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica*, **32**, 122–136.

Prelec, D. (1998). The probability weighting function. *Econometrica*, **66**, 497–527.

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior and Organization*, **3**, 323–343.

Richardson, J. (1994). Cost utility analysis: What should be measured?. *Social Science and Medicine*, **36**, 7–20.

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica*, **57**, 571–587.

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

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

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, **76**, 31–48.

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

Tversky, A., Sattah, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review*, **95**, 371–384.

Tversky, A., & Wakker, P. P. (1995). Risk attitudes and decision weights. *Econometrica*, **63**, 1255–1280.

Varey, C., & Kahneman, D. (1992). Experiences extended across time: Evaluation of moments and episodes. *Journal of Behavioral Decision Making*, **5**, 169–186.

Viscusi, W. K., & Moore, M. J. (1989). Rates of time preference and valuations of the duration of life. *Journal of Public Economics*, **38**, 297–317.

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

Weber, E. U. (1994). From subjective probabilities to decision weights: The effect of asymmetric loss functions on the evaluation of uncertain outcomes and events. *Psychological Bulletin*, **115**, 228–242.

Weber, E. U., & Kirsner, B. (1997). Reasons for rank-dependent utility evaluation. *Journal of Risk and Uncertainty*, **14**, 41–61.

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

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

Received: November 19, 1998