Journal of Health Economics 28 (2009) 1039–1047

Contents lists available at ScienceDirect

Journal of Health Economics

journal homepage: www.elsevier.com/locate/econbase

The predictive validity of prospect theory versus expected utility in health utility measurement

Jose Maria Abellan-Perpiñan[a], Han Bleichrodt[b,*], Jose Luis Pinto-Prades[c]

[a] *Department of Applied Economics, University of Murcia, Murcia, Spain*
[b] *Erasmus School of Economics & iMTA/iBMG, Erasmus University Rotterdam, Rotterdam, Netherlands*
[c] *Department of Economics, University Pablo d'Olavide & Fundación Centro de Estudios Andaluces, Andaluces, Sevilla, Spain*

ARTICLE INFO

Article history:
Received 20 November 2007
Received in revised form 23 July 2009
Accepted 10 September 2009
Available online 20 September 2009

JEL classification:
D81
I10

Keywords:
Health utility measurement
Prospect theory
Expected utility
Standard gamble
Time trade-off

ABSTRACT

Most health care evaluations today still assume expected utility even though the descriptive deficiencies of expected utility are well known. Prospect theory is the dominant descriptive alternative for expected utility. This paper tests whether prospect theory leads to better health evaluations than expected utility. The approach is purely descriptive: we explore how simple measurements together with prospect theory and expected utility predict choices and rankings between more complex stimuli. For decisions involving risk prospect theory is significantly more consistent with rankings and choices than expected utility. This conclusion no longer holds when we use prospect theory utilities and expected utilities to predict intertemporal decisions. The latter finding cautions against the common assumption in health economics that health state utilities are transferable across decision contexts. Our results suggest that the standard gamble and algorithms based on, should not be used to value health.

1. Introduction

A crucial ingredient in health care evaluation is measuring the utility of health states. There is much ambiguity in the existing measurement methods with different methods that have the same prima facie plausibility leading to systematically different results. A danger of this divergence in measured utilities is that health care evaluations come to depend on the measurement method used. This would not be a cause of concern if it were known which method should be preferred. Unfortunately, no consensus exists on this question.

Most health care evaluations today still assume expected utility. For example, the standard gamble (SG), a widely used method for measuring health utility, is based on expected utility. Expected utility was the leading normative and descriptive theory of decision making in the 1980s when economic evaluations of health care took off. Since then, many descriptive deficiencies of expected utility have been documented (Starmer, 2000). It is now widely accepted that expected utility is not valid as a descriptive theory of human decision making. Evidence of violations of expected utility for health outcomes includes Llewellyn-Thomas et al. (1982), Rutten-van Mölken et al. (1995), and Stalmeier and Bezembinder (1999). Bleichrodt (2002) argued that the violations of expected utility imply that the SG overestimates the utility of health. Bleichrodt et al. (2007) provided empirical evidence to sustain this argument.

The descriptive deficiencies of expected utility have led to a multitude of alternative theories of decision under risk. Of these theories, prospect theory is currently the most important. Because prospect theory is descriptively superior to expected utility, it can lead to better health evaluations. Bleichrodt et al. (2001) proposed adjustments of existing health valuation methods under risk, in particular the SG, based on prospect theory. They showed that their adjustments lead to higher internal consistency in the sense that theoretically equivalent methods give similar results. Further support for their adjustments is in van Osch et al. (2004), van Osch et al. (2006), and Bleichrodt et al. (2007). None of these studies examined the external validity of prospect theory, i.e. whether the adjustments are better able to predict people's choices than the traditional methods based on expected utility. The studies exploring internal consistency were all based on a common response mode, typically matching. It is possible that the adjustments correct for a

* Corresponding author.
E-mail addresses: dionisos@um.es (J.M. Abellan-Perpiñan), bleichrodt@ese.eur.nl (H. Bleichrodt), jlpinpra@upo.es (J.L. Pinto-Prades).

0167-6296/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhealeco.2009.09.002

heuristic implied in the common response mode. They could work "as if" people are prospect theory maximizers while in reality they are not. This explanation can be ruled out when preferences are elicited through different response modes, as in tests of external validity.

External validity is the topic of this paper. We investigate the potential of prospect theory to lead to better health evaluations. We evaluate risky prospects under prospect theory and under expected utility and examine to what extent these evaluations are consistent with directly elicited rankings and choices. The purpose of the paper is purely descriptive. We do not take ranking and choice as gold standards but explore how some simple measurements together with expected utility and prospect theory predict rankings and choices between more complex stimuli. Even though we do not take rankings and choices as gold standards, we do believe that they are to some extent reflective of people's preferences.

This exploration will give insight into which theory performs better in describing health decisions under risk. In health economics it is common practice to assume that utility is not only applicable within one decision context, e.g. risk, but that it is transferable across decision contexts. For example, SG utilities, measured under risk, are routinely used in societal decisions about the allocation of health care resources, i.e. in welfare evaluations. Likewise, time trade-off utilities, measured in an intertemporal setting, are used both in decisions under risk and in welfare evaluations.

Given this common practice, we also compared the performance of the evaluations based on prospect theory and on expected utility in predicting intertemporal choices and rankings. We not only compared expected utility and prospect theory but also compared them with evaluations based on the time trade-off. Bleichrodt and Johannesson (1997) explored the external validity of the SG in an intertemporal ranking task and found that it was significantly worse than that of the time trade-off (TTO). Their finding suggests problems with the transferability of utility across decision contexts. This paper provides more extensive evidence by comparing SG, the SG adjusted for the violations of expected utility modeled by prospect theory, and TTO both in risky and in intertemporal decisions. The results provide more insight into the question whether there is one measure of health utility that can be used in all types of health evaluations or whether different evaluations call for different measures.

At this point the reader may wonder why the results for the adjusted SG can be different from the results for the SG. Choice and ranking only give ordinal information about preferences and the adjusted SG is a monotonic transformation of the SG. The answer is that SG and adjusted SG are not used straightforwardly to derive choices, but are used in an additive evaluation of nonconstant profiles. Hence, they are used in a cardinal sense and this is why they can give different results.

The structure of the paper is as follows. Section 2 provides background. Section 3 describes the experiment we used and Section 4 its results. Section 5 discusses our main findings and Section 6 concludes the paper.

2. Background

Let $q = (q_1, \ldots, q_T)$ denote a *health profile* that yields health state q_t in period t. T is the last period of the decision maker's life. A health profile is *constant* if $q_t = Q$ for all t. For notational convenience, constant health profiles will be written as (Q, T), denoting T years in health state Q. By q_pq' we denote the *prospect* that gives health profile q with probability p and health profile q' with probability $1 - p$. If $q = q'$ or $p = 0$ or $p = 1$ the prospect is *riskless*, otherwise it is *risky*. By $\succcurlyeq$ we denote the preference relation "at least as good as" defined over prospects. Strict preference is denoted by $\succ$ and indifference by $\sim$. By restricting attention to riskless prospects, $\succcurlyeq$ defines a preference relation over health profiles. It is implicit in the notation q_pq' that q is at least as good as q': $q \succcurlyeq q'$.

Expected utility holds if prospects q_pq' are evaluated by $pU(q) + (1 - p)U(q')$ and preferences and choices correspond with this evaluation. U is a utility function over constant health states that is unique up to unit and location.

Prospect theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992) generalizes expected utility in two ways. First, prospect theory does not assume that preferences are linear in probability but allows for *probability weighting*. Probability weighting is modeled through a *probability weighting function*, which is increasing and assigns weight 0 to probability 0 and weight 1 to probability 1. The second deviation from expected utility modeled by prospect theory is *sign-dependence*: people perceive outcomes as gains and losses with respect to a *reference point* r. A *gain* is strictly preferred to the reference point and a *loss* strictly less preferred than the reference point. People are assumed to be more sensitive to losses than to absolutely commensurate gains, a phenomenon known as *loss aversion*. Sign-dependence also affects the weighting of probabilities: prospect theory allows that probability weighting for gains w^+ is different from probability weighting for losses w^-.

Under prospect theory, prospects q_pq' involving no losses are evaluated as $U(r) + w^+(p)\ (U(q) - U(r)) + (1 - w^+(p))\ (U(q') - U(r))$, prospects involving no gains are evaluated as $U(r) - \lambda w^-(1 - p)$ $(U(r) - U(q')) - \lambda(1 - w^-(1 - p))\ (U(r) - U(q))$ and *mixed prospects*, prospects involving both a gain and a loss, are evaluated as $U(r) + w^+(p)\ (U(q) - U(r)) - \lambda w^-(1 - p)\ (U(r) - U(q'))$. In these formulas U is a utility function over health profiles that is unique up to unit and location and λ is a coefficient that reflects loss aversion. In the literature there exist several definitions of loss aversion and λ can have different meanings depending on the definition used. For a discussion of the various definitions of loss aversion see Köbberling and Wakker (2005) and Abdellaoui et al. (2007b).

In the *standard gamble*, the probability p is elicited such that a decision maker is indifferent between (Q,T) for sure and a risky prospect $(\mathrm{FH},T)_p\mathrm{Death}$, where FH denotes full health. Under expected utility, this indifference implies that

$$U(Q, T) = pU(\mathrm{FH}, T) + (1 - p)U(\mathrm{Death}). \tag{1}$$

Under prospect theory the evaluation depends on the reference point. Several studies have provided evidence that people evaluate SG questions by taking the sure outcome (Q,T) as their reference point (Morrison, 2000; Bleichrodt et al., 2001; Robinson et al., 2001; van Osch et al., 2004, 2006). This implies that the risky prospect $(\mathrm{FH},T)_p\mathrm{Death}$ is *mixed*, i.e. it yields a gain, an improvement in health from (Q,T) to (FH,T) with probability p and a loss, a deterioration in health from (Q,T) to Death, with probability $1 - p$. The evaluation of the SG question is then equal to

$$\begin{aligned} &U(Q, T) + w^+(p)(U(\mathrm{FH}, T) - U(Q, T)) \\ &\quad - \lambda w^-(1 - p)(U(Q, T) - U(\mathrm{Death})). \end{aligned} \tag{2}$$

The term $U(Q,T)$ appears in (2) because it is the reference point. The outcomes of the standard gamble are evaluated as deviations from the reference point.

Throughout this paper we will assume that the utility function over health profiles is equal to

$$U(q_1, \ldots, q_T) = \sum_{t=1}^{T} \rho_t H(q_t), \tag{3}$$

Table 1
The description of health states A and B.

Health state A	Health state B
Some problems walking about	Some problems walking about
Some problems performing self-care activities (e.g. eating, washing, dressing)	Some problems performing self-care activities (e.g. eating, washing, dressing)
No problems performing usual activities (e.g. work, study, family or leisure activities)	Unable to perform usual activities (e.g. work, study, family or leisure activities)
Moderate pain or discomfort	Moderate pain or discomfort
Moderately anxious or depressed	Moderately anxious or depressed

where ρ_t is a discount weight that specifies the weight given to period t and H is a utility function over health status. We will refer to Eq. (3) as the *discounted QALY model*. Special cases of Eq. (3) are the *undiscounted QALY model*, for which $\rho_t = 1$ for all t, and the QALY model with *constant discounting*, for which $\rho_t = 1/(1+r)^{t-1}$ for all t. Both special cases have been widely applied in health care evaluations.

Define $L(s) = \sum_{t=1}^{s} \rho_t$. L can be interpreted as a utility function over life duration. Eq. (3) implies that U in Eqs. (1) and (2) is equal to $U(Q,T) = H(Q) \times L(T)$. Preference conditions for the discounted QALY model have been given by Bleichrodt and Quiggin (1997) for variable health profiles and by Miyamoto et al. (1998) for constant health profiles. The discounted QALY model entails that more years in full health are always desirable, an implication used in what follows.

Under the discounted QALY model and the common scaling $H(\mathrm{FH}) = 1$ and $U(\mathrm{Death}) = 0$, Eq. (1) yields $H(Q) = p$. This is the common way in which the standard gamble is evaluated in health economics. Under prospect theory this evaluation is no longer correct. Applying Eq. (2) and some elementary algebra, we obtain

$$H(Q) = \frac{w^{+}(p)}{w^{+}(p) + \lambda w^{-}(1-p)}. \qquad (4)$$

To compute $H(Q)$, the probability weighting functions and the loss aversion coefficient must be known. We will assume the estimates obtained by Tversky and Kahneman (1992). Bleichrodt et al. (2001) found that these estimates performed well at the aggregate level. Empirical studies that estimated probability weighting and loss aversion parameters generally obtained results that were close to Tversky and Kahneman's estimates (Gonzalez and Wu, 1999; Abdellaoui, 2000; Bleichrodt and Pinto, 2000; Abdellaoui et al., 2007a). Adopting Tversky and Kahneman's (1992) estimates implies that health state utilities $H(Q)$ can be computed from the responses to SG questions by using Table 1 in Bleichrodt et al. (2001). We will refer to the utilities thus obtained as the *adjusted SG utilities.*

The TTO elicits the number of years T^* in full health that makes a decision maker indifferent to T years in some impaired health state Q. Under the discounted QALY model and the scaling $H(\mathrm{FH}) = 1$ this implies that $H(Q) = L(T^*)/L(T)$. L is generally assumed linear in the empirical literature on the TTO. Then $H(Q) = T^*/T$.

In the above derivations we used the same utility function H in the evaluation of the SG and the TTO. Several authors have argued that this is not allowed because utility is context-specific and there is no unifying concept of utility. The utility function elicited under risk may be different from the function elicited in the intertemporal context employed in the TTO (Arrow, 1951, pp. 425; Dyer and Sarin, 1982; Fishburn, 1989; Gafni et al., 1993). Others have argued in favor of the existence of one unifying concept of utility (Harsanyi, 1955; Richardson, 1994; Wakker, 1994). Empirical evidence that utility in different decision contexts is similar was obtained by Stalmeier and Bezembinder (1999) for health outcomes and by Abdellaoui et al. (2007a) for money outcomes. As mentioned before, in health economics it is common to use the same utilities in different decision contexts. Our empirical results will shed additional light on the question whether this is justified.

3. Experiment

3.1. General idea

We computed the number of QALYs of 10 health alternatives, 5 risky prospects and 5 intertemporal health profiles, based on EU, PT, and the TTO and compared the implied ranking of the health alternatives with the directly elicited ranking and with the ranking implied by directly observed choices.

3.2. Subjects

Subjects were sixty-five ($N = 65$) economics students (aged between 22 and 29) from the University of Murcia. They were paid €36 to participate in five experimental sessions. In this paper, we only use the results from the first, second, and fifth session. Each experimental session lasted approximately one hour. The experiment was carried out in small group sessions with at most six subjects per session. The sessions were separated by at least one week. Prior to the actual experiment, the questionnaire was tested in several pilot sessions.

3.3. Stimuli

We elicited the utility of two EQ-5D health states, 22,122 and 22,322. The description of the health states is given in Table 1. These health states have Spanish EuroQol values 0.596 and 0.110. We chose these health states as we wanted two states worse than full health for which the preference ordering was obvious. Throughout the experiment, the health states were labeled A and B. Health state A dominates health state B in the sense that it yields a level of functioning that on each dimension is at least as good as the corresponding level of B. Full health was described as no limitations on any of the dimensions.

We asked six questions both for the SG[1] and for the TTO by combining the two health states with three different values for T: 13 years, 24 years, and 38 years. We used lower durations than subjects' life expectancy to avoid perception problems. We learnt from the pilot sessions that subjects found it hard to perceive living for longer than their life expectancy.

Table 2 displays the health alternatives used. We selected five risky prospects and five intertemporal profiles. The notation 9A + 4B + 4FH stands for nine years in health state A followed by 4 years in health state B followed by 4 years in full health. We opted for health alternatives involving different probabilities and

[1] Recall that the adjusted SG can be computed from the response to the SG question. No additional questions are needed to compute it.

Table 2
The health alternatives used.

Risky prospects
$11A_{0.63}$Death
$17A_{0.5}$Death
$11B_{0.5}7A$
$6B_{0.5}6FH$
$14A_{0.45}7B$
Intertemporal profiles
14A + 3FH
9A + 4B + 4FH
4FH + 13B
1FH + 13B + 3FH
2FH + 4A + 8B

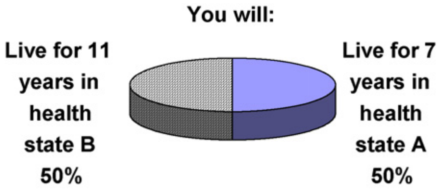

Fig. 1. Display of the risky health prospects.

different life durations for which the preference ordering was not immediately obvious. In the risky prospects we only used health profiles of constant quality to keep the tasks tractable. All health profiles ended with death. The health alternatives were printed on cards. The risky health prospects were displayed as pie charts with the size of each pie corresponding to the size of the probability. Fig. 1 gives an example. The intertemporal health profiles were displayed as stacked bars with the size of each component of the bar corresponding to the duration of the health state. Fig. 2 gives an example.

3.4. Procedures

Preferences in the SG and TTO questions were elicited through a choice-based procedure in which only the parameter that we sought to elicit varied. We always started with parameter values for which one of the alternatives was clearly better than the other and then zeroed in on the parameter value for which subjects were indifferent between the alternatives.

Recruitment of subjects took place one week before the first session of the actual experiment started. At recruitment, subjects received information about the experiment and they were asked to read the descriptions of the two health states. In addition, the subjects were handed a practice question for the SG method. They were asked to answer this practice question at home. This procedure intended to familiarize subjects with the SG method. Prior to the start of the first experimental session, during which the SG method was administered, the subjects were asked to explain their answer to the practice question. When we were not convinced that a subject understood the task, we explained it again until we were convinced that the task was understood. The same procedure was used for the TTO. The subjects received a practice question to take home showing the method that would be administered in the next session, and they had to explain their answer to this question before the actual experiment started.

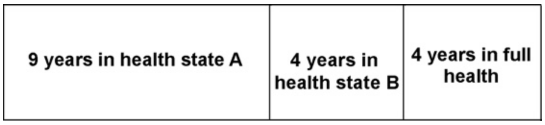

Fig. 2. Display of the intertemporal health profiles.

At the beginning of each experimental session, instructions were read aloud and an additional practice question was given. The order in which the methods were administered was: first session SG and ranking, second session TTO, and fifth session choices. The experiment was part of a larger experiment. The presence of the other experimental tasks and the delay of at least one week between the sessions made it unlikely that the subjects would recall their previous answers or would note the relationship between the sessions. To avoid order effects, we varied the order in which the different questions were asked within a session. To reduce response errors, subjects had to confirm the elicited indifference value after each SG or TTO question. The final comparison was shown again and subjects were asked whether they agreed that the displayed options were equivalent. If not, the elicitation procedure for that question was started anew.

In the ranking task, subjects were given the cards with the health alternatives in arbitrary order and were asked to rank these on the table in front of them. In the choice questions, the cards were pitted against each other in arbitrary order and subjects were asked which alternative they preferred. To reduce the cognitive burden, we only asked subjects to compare risky prospects with risky prospects and intertemporal profiles with intertemporal profiles in choice and ranking. Hence, they did not compare risky prospects with intertemporal profiles.

3.5. Analysis

Adjusted SG utilities were computed from the responses to the SG questions using Table 1 in Bleichrodt et al. (2001). Once we had determined the health state utilities for each of the three methods (SG, adjusted SG, and TTO) and for each of the three values of T (13 years, 24 years, and 38 years) we used these to compute for each subject the number of QALYs for each of the health alternatives.

The number of QALYs of the risky health prospects was computed both by applying expected utility and by applying prospect theory. Under EU, the number of QALYs of the risky health prospect $14A_{0.45}7B$ using the data for $T = 38$ years was, for example,

$$0.45 \times L(14) \times H_{\mathrm{SG},38}(\mathrm{A}) + 0.55 \times L(7) \times H_{\mathrm{SG},38}(\mathrm{B}), \tag{5}$$

where $H_{\mathrm{SG},38}$ denotes the standard gamble utility measured with $T = 38$ years.

Under prospect theory the number of QALYs of the risky health prospect $14A_{0.45}7B$ was

$$w(0.45) \times L(14) \times H_{\mathrm{PT},38}(\mathrm{A}) + (1 - w(0.45)) \times L(7) \times H_{\mathrm{PT},38}(\mathrm{B}), \tag{6}$$

where $H_{\mathrm{PT},38}$ denotes the adjusted standard gamble utility determined with $T = 38$ years. When using the TTO utilities, it is not a priori clear whether we should use prospect theory or expected utility to compute the number of QALYs. Hence, we used both. That is, we computed the number of QALYs both according to (5), with $H_{\mathrm{TTO},38}$ instead of $H_{\mathrm{SG},38}$, and according to (6), with $H_{\mathrm{TTO},38}$ instead of $H_{\mathrm{PT},38}$.

There is a difficulty in applying prospect theory. To use prospect theory, it is crucial to know the ordering of the profiles involved in a prospect. For most prospects this caused no problems, but for the prospect $11B_{0.5}7A$ it is not clear what the preference ordering of 11B and 7A is. We computed the preference implied by $H_{\mathrm{PT},\,j}$, $j = 13$, 24, 38 years, and the discounted QALY model and applied prospect theory accordingly. For example, if $H_{\mathrm{PT},13}(\mathrm{A}) = 0.7$ and $H_{\mathrm{PT},13}(\mathrm{B}) = 0.4$ and there is no discounting, i.e. L is linear, then the QALY model predicts that 7A is preferred to 11B but if $H_{\mathrm{PT},13}(\mathrm{A}) = 0.6$ and $H_{\mathrm{PT},13}(\mathrm{B}) = 0.4$ then 11B is predicted to be preferred to 7A. The example shows that the ordering of 11B and 7A can vary across subjects.

The number of QALYs for the intertemporal profiles was computed by applying Eq. (3). For example, the number of QALYs of the profile 9A + 4B + 4FH according to expected utility and using the data for $T = 24$ years is equal to

$$L(9)H_{\mathrm{SG},24}(\mathrm{A}) + (L(13) - L(9))H_{\mathrm{SG},24}(\mathrm{B}) + (L(17) - L(13)). \quad (7)$$

To compute Eqs. (5)–(7) we must know L. We first assumed that L was linear, i.e. we assumed the undiscounted QALY model. The reason to perform this analysis was the widespread use of the undiscounted QALY model in health economics. The number of undiscounted QALYs according to EU, PT, and TTO implied a ranking of the risky health prospects and of the intertemporal health profiles. This ranking could be compared with the directly elicited ranking. The elicited pairwise choices also led to a rank ordering of the risky health prospects and of the intertemporal health profiles in the sense that the rank of a health alternative was determined by the number of times it was chosen over another health alternative. In the case of ties both health alternatives were assigned the same rank. The consistency of the three methods with the directly elicited ranking and the ranking implied by the observed choices was assessed by the Spearman rank correlation coefficient. All results were similar when Kendall's tau was used and these results are therefore not reported.

To reduce the possibility that our results were driven by violations of the undiscounted QALY model, we also analyzed the data allowing for curvature of the utility for life duration. We adopted an exponential specification for L. The *exponential family* corresponds to constant discounting and is defined by $L(s) = (e^{rs} - 1)/(e^{r} - 1)$ if $r \neq 0$ and by $L(s) = s$ if $r = 0$. The utility for life duration is concave if $r < 0$ and convex if $r > 0$. To allow for deviations from constant discounting (van der Pol and Cairns, 2002), we also evaluated the data under a power specification for L. We estimated the exponential and power families by nonlinear least squares. The results were similar, but convergence was better for the exponential specification. Hence, we will report the exponential results in Section 4.

All reported statistical tests are paired t-tests with unequal variances. We also performed nonparametric tests but these led to the same results and are not reported.

4. Results

The QALY model implies that preferences should satisfy *monotonicity*: for a given health state Q, additional life-years are always desirable (for all T,T', if $T > T'$ then $(Q,T) \succ (Q,T')$ or they are always undesirable (for all T,T', if $T > T'$ then $(Q,T) \prec (Q,T')$. Violations of the QALY model, e.g. preferences implying that additional life-years are initially considered desirable but undesirable after some threshold, the so-called maximal endurable time first reported by Sutherland et al. (1982), cannot be accommodated by the QALY model as explained in detail elsewhere (Miyamoto et al., 1998). The responses of 19 of our 65 subjects violated monotonicity and these subjects were excluded. Hence, the final analyses included the responses from 46 subjects. Our motivation to exclude the subjects violating monotonicity was the following. SG and TTO are methods that are used within the QALY model. They derive their relevance within the context of the QALY model. If the QALY model does not hold then the question which method to use becomes irrelevant.

As expected, violations of monotonicity occurred primarily for the less attractive health state B. Only two subjects violated monotonicity for health state A. The subjects violating monotonicity exhibit a preference reversal as introduced by Stalmeier et al. (1997). The seventeen subjects who violated monotonicity only for health state B had similar SG and adjusted SG valuations for health state A as the subjects in our analysis. Their TTO valuations were, however, significantly lower.

Fig. 3 shows the mean utilities of health states A and B according to the SG, adjusted SG, and TTO. The medians were similar.[2] The utility of health state A was significantly higher ($p < 0.01$ in all comparisons) than that of health state B according to all three methods as expected given that A is a better health state than B. The figure displays a dichotomy between the methods: SG and TTO are close but differ substantially from the adjusted SG, the difference being around 0.25. The difference between SG and TTO on the one hand and adjusted SG on the other hand is significant ($p < 0.001$). SG and TTO differ significantly for health state A and durations 13 years and 38 years ($p < 0.01$ in both cases) and for health state B and duration 24 years ($p = 0.023$). In the other three comparisons the differences are not significant ($p > 0.10$).

Recall that the EuroQol valuations for A and B are 0.596 and 0.110. Our valuations for A are significantly higher for SG and TTO (with the exception of the SG and $T = 13$ years) and significantly lower based on the adjusted SG ($p < 0.001$). For health state B the valuations are significantly higher regardless of the method used ($p < 0.001$).

4.1. Risky prospects

Table 3 shows the mean ranks of the five risky prospects for the ranking and the choice data and for EU, PT, and the TTO. The data for the TTO are based both on expected utility, $\mathrm{TTO}_{\mathrm{EU}}$, and on prospect theory, $\mathrm{TTO}_{\mathrm{PT}}$, i.e. Eqs. (5) and (6) with $H_{\mathrm{TTO},j}$. Lower numbers reflect more attractive prospects, i.e. rank 1 is the best score and 5 the worst. The ranking and choice data are comparable, except that prospects $6\mathrm{B}_{0.5}6\mathrm{FH}$ and $17\mathrm{A}_{0.5}\mathrm{Death}$ are considered significantly more attractive in choice than in ranking ($p < 0.001$ and $p = 0.018$ respectively).

Prospects involving the possibility of immediate death are considered the least attractive. The preference of prospect $11\mathrm{A}_{0.63}\mathrm{Death}$ over prospect $17\mathrm{A}_{0.5}\mathrm{Death}$ indicates that subjects took differences between probabilities into account. The expressed rankings and choices suggest that subjects tried to minimize the risk of death. The prospect $6\mathrm{B}_{0.5}6\mathrm{FH}$ is considered relatively attractive even though it involves only short life durations. This observation suggests that subjects were sensitive to differences in quality of life. It also suggests that subjects did not value additional life duration as much as the undiscounted QALY model implies. This finding is unlikely to be caused by considerations of maximal endurable time (Sutherland et al., 1982) as the subjects who violated monotonicity with respect to life duration were excluded from the analyses.

Table 3 shows that the evaluation under prospect theory is more consistent with the rankings and choices than the evaluation under expected utility. This pattern is clearer when the Spearman rank correlation coefficients are considered. Fig. 4 illustrates. It shows the mean Spearman rank correlation coefficients for the evaluations under expected utility and under prospect theory. Part A displays the results for the ranking data, part B for the choice data. It is immediately obvious from the figure that prospect theory is substantially and significantly more consistent with subjects' directly

[2] The standard deviations of the different utilities were similar for the three durations that we used in the elicitations. Mean standard deviations for health states A and B were 0.129 and 0.098 for the SG, 0.078 and 0.062 for the adjusted SG, and 0.105 and 0.078 for the TTO.

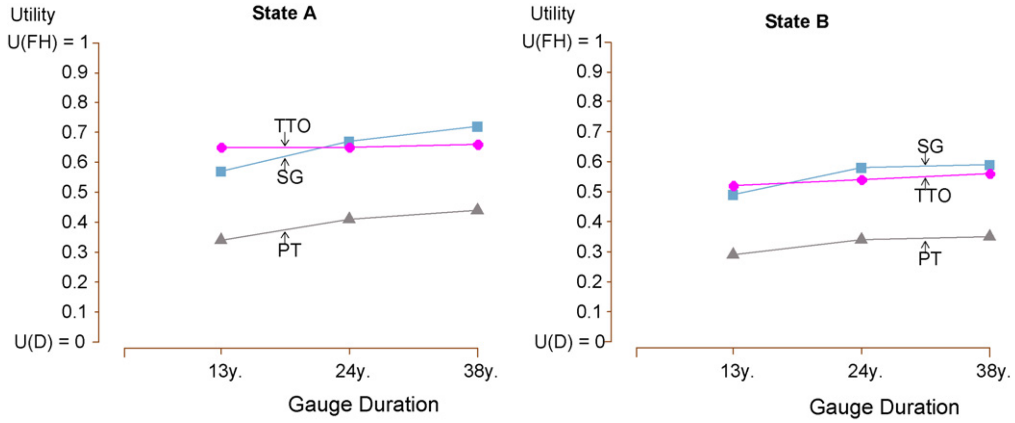

Fig. 3. Mean utilities for health states A and B according to the three methods.

Table 3
Mean ranks of the risky prospects according to the three methods.

Prospect	Ranking	Choice	EU	PT	TTO_{EU}	TTO_{PT}
$14A_{0.45}7B$	1.48	1.52	1.13	1.55	1.09	1.03
$6B_{0.5}6FH$	2.52	1.98	4.33	1.57	4.23	3.72
$11B_{0.5}7A$	2.61	2.83	2.78	3.16	2.98	2.30
$11A_{0.63}$Death	3.78	3.50	4.37	4.99	4.36	4.90
$17A_{0.5}$Death	4.61	4.22	2.40	3.72	2.35	3.06

elicited preferences than expected utility ($p < 0.001$). This holds both for the ranking and for the choice data, but the difference is particularly large for the choice data. Our finding that violations of expected utility are more pronounced in choice than in ranking is consistent with Bateman et al. (2007).

Fig. 5 shows the rank correlation coefficients for the evaluation under expected utility and under prospect theory when we use the TTO utilities. The figure shows that the consistency with choices and ranking improves considerably when we use prospect theory instead of expected utility. A comparison between Figs. 4 and 5 shows that consistency with the directly elicited rankings and choices, is lower when the TTO is used than when the adjusted SG is used. The difference is significant ($p = 0.008$ for rankings and $p < 0.001$ for choices).

The above conclusions are also confirmed when we look at the individual level data. Overall prospect theory leads to the highest rank correlation coefficients for 47.8% of the subjects, expected utility for 13.0% of the subjects, and the TTO (based on prospect theory, Eq. (6) for 21.7% of the subjects (for the remaining subjects there is no clear winner).

There is a lot of variation at the individual level but the spread in the rank correlation coefficients is comparable across the methods. The mean standard deviations of the Spearman rank correlation coefficients are 0.38 for prospect theory, 0.44 for expected utility, and 0.36 for the TTO.

4.2. *Intertemporal profiles*

Table 4 shows the mean rank of the five intertemporal profiles based on the ranking data, the choice data, and on QALYs computed under EU, PT, and TTO. The data on rankings and choices are to a large extent comparable except that profiles 14A + 3FH and 1FH + 13B + 3FH are significantly more attractive in the choice data ($p = 0.015$ and $p = 0.042$ respectively). The mean rank of the profile 4FH + 13B is almost the same as the mean rank of the profile 1FH + 13B + 3FH in the ranking data. This suggests zero time preference at the aggregate level. In the choice data 1FH + 13B + 3FH is considered significantly ($p = 0.001$) more attractive than 4FH + 13B suggesting negative discounting, i.e. convex utility for life duration.

Several studies in decision theory found evidence that splitting alternatives into more detailed levels increases their attractiveness (e.g. Weber et al., 1988). This effect did not influence our findings. If it were present profiles consisting of three levels, 9A + 4B + 4FH, 2FH + 4A + 8B, and 1FH + 13B + 3FH, should be considered relatively attractive in choice and ranking compared with the predicted rank-

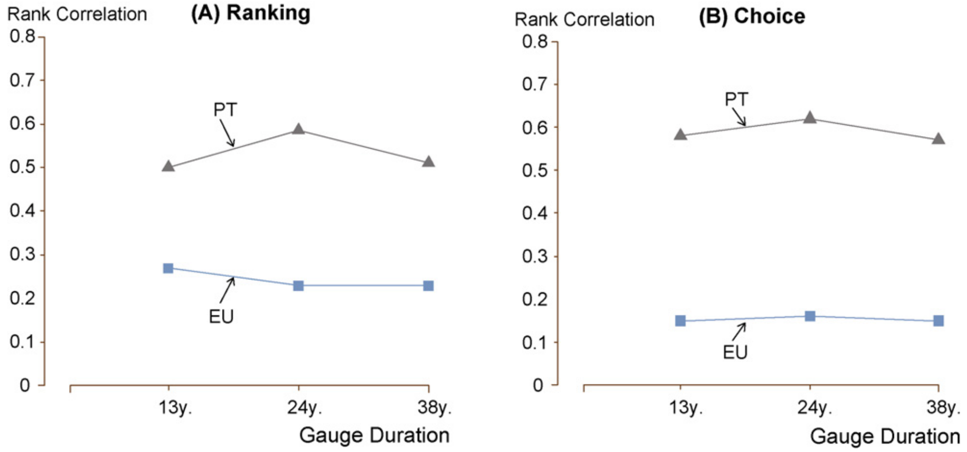

Fig. 4. Mean Spearman rank correlations for the risky prospects based on the SG responses.

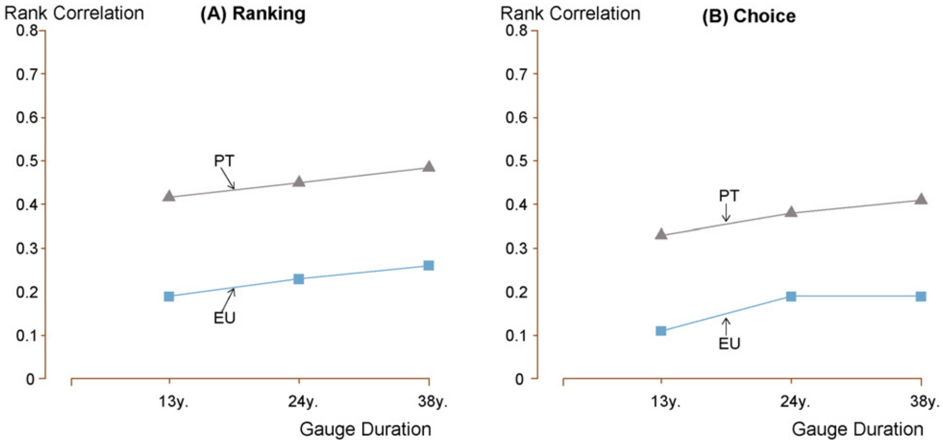

Fig. 5. Mean Spearman rank correlations for the risky prospects based on the TTO responses.

Table 4
Mean ranks of the intertemporal profiles according to the three methods.

Profile	Ranking	Choice	EU	PT	TTO
14A + 3FH	1.35	1.04	1.93	2.54	1.80
9A + 4B + 4FH	2.04	1.96	2.01	1.65	2.00
2FH + 4A + 8B	3.57	3.57	4.96	4.96	4.99
4FH + 13B	4.00	4.28	2.51	2.41	2.58
1FH + 13B + 3FH	4.04	3.78	2.51	2.41	2.58

ing by EU, PT, and TTO. This is true for 2FH + 4A + 8B but not for the other two profiles. To the contrary, the profile 1FH + 13B + 3FH is considered relatively unattractive in choice and ranking.

Fig. 6 displays the mean Spearman rank correlation coefficients for the intertemporal profiles when we assume that the undiscounted QALY model holds. Fig. 6A shows that undiscounted QALYs based on the TTO are most consistent with the ranking data. This finding is in line with the main conclusion in Bleichrodt and Johannesson (1997). The difference between TTO and EU, i.e. the SG, is, however, less pronounced than in Bleichrodt and Johannesson (1997) and does not reach statistical significance ($p > 0.10$ in all three comparisons). TTO and SG are more consistent with the ranking data than PT, the adjusted SG. The differences between TTO and PT are, however, only significant for $T = 13$ years ($p = 0.028$, $p > 0.10$ in all other comparisons).

The choice data are comparable. The TTO is not significantly more consistent with the choice data than the SG ($p > 0.10$ in all cases), and even though TTO and SG are more consistent with the choice data than PT, the difference between TTO and PT is only significant for $T = 13$ ($p = 0.001$). The difference is small and insignificant for durations 24 years and 38 years.

The individual data analysis mirrors the aggregate analysis. For 26.1% of the subjects the TTO yields the highest rank correlation coefficients, for 17.4% the SG, and for 4.3% PT. For most of the large proportion of unclassified subjects TTO and SG perform equally well. The variation in the individual rank correlation coefficients is larger than for the risky prospects, but, again, comparable across the methods. The mean standard deviations of the Spearman rank correlation coefficients are 0.50 for the TTO, 0.52 for PT, and 0.55 for the SG.

4.3. Discounted QALYs

So far we have assumed that the utility for life duration is linear. Several studies have indicated that subjects do not have linear but concave utility for life duration (e.g. Stiggelbout et al., 1994; Bleichrodt and Pinto, 2005). Our data on the risky prospects also suggest that the assumption of linear utility for life duration may be inappropriate. One way of incorporating concave utility is by discounting QALYs. In the literature QALYs are commonly discounted by 3% or 5%. Applying these discount rates does not affect the conclusions. Fig. 7 shows the results for the choice data when 5% discounting is applied. The results for the ranking data and for 3% discounting are similar. In general, discounting tends

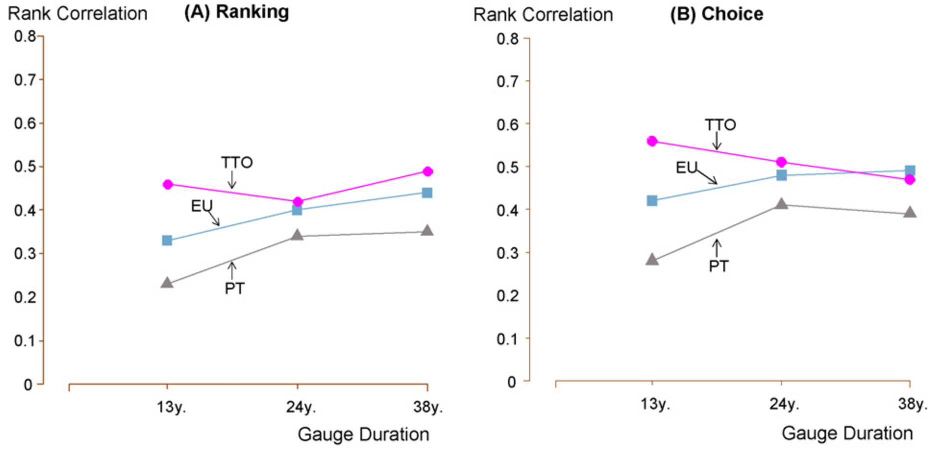

Fig. 6. Mean Spearman rank correlations for the intertemporal profiles.

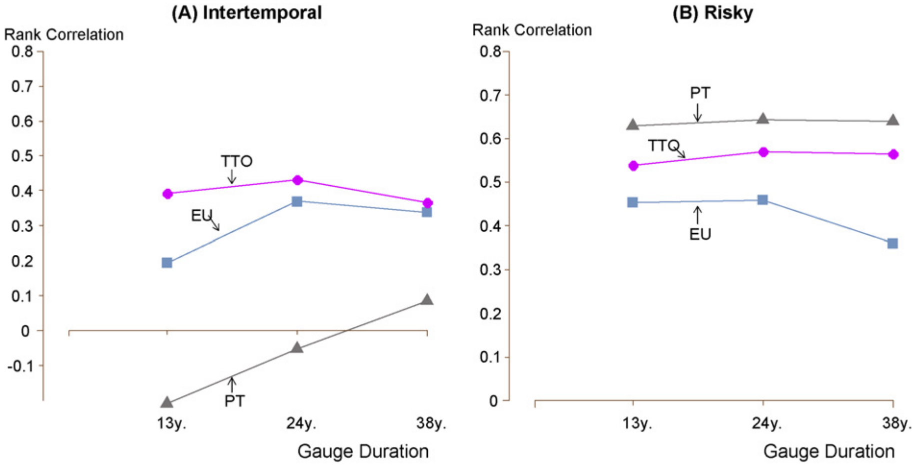

Fig. 7. Mean Spearman rank correlation coefficients for the choice data with 5% discounting.

to reduce the rank correlation coefficients for the intertemporal data and increases the rank correlation coefficients for the risky data.

The above observation is confirmed when we estimate optimal discounting parameters for each subject. For the intertemporal data close to zero discounting is optimal. For the risky data the optimal parameters indicate strong discounting (>30%).

The large difference in optimal discounting between the risk and the intertemporal domain does not necessarily imply that discounting is very different across the two domains. It is plausible that discounting corrects for a difference in decision rules across the two domains. For instance, as we noticed before, subjects tended to minimize the risk of death in the evaluation of the risky prospects. As a consequence, the prospect $17A_{0.5}$Death was considered relatively unattractive. But this was also the prospect with the longest life duration. To make this prospect unattractive life-years had to be discounted at a high rate. High discounting in this case mimics the decision rule to minimize the risk of death.

5. Discussion

This paper has explored whether prospect theory leads to better health evaluations than expected utility. The answer is clearly yes when we consider preferences over risky prospects. The consistency of prospect theory with directly elicited choices and rankings is much higher than that of expected utility.

The second question explored in this paper was to what extent utilities elicited in one particular decision context can be transferred to another decision context. The two contexts that we studied were risky decisions and intertemporal decisions. Our answer here is more ambiguous. For risky decisions, the adjusted SG was most consistent with choices and rankings. The differences with TTO and SG were significant. For intertemporal decisions, TTO utilities were most consistent with choices and rankings, but the difference with the SG was not significant and with the adjusted SG it was only significant for duration 13 years. For the other two durations it was not significant.

Transferability of utility is commonly assumed in health economics. Our findings caution against routinely taking such transferability for granted. Of course, we do not claim to provide the definitive answer on the transferability of utility. A caveat that should be made is that we assumed the discounted QALY model throughout. We took care of some violations of this model by excluding subjects who did not always prefer more life-years to less. However, other violations of the QALY model are conceivable. The assumption of intertemporal additivity seems particularly problematic. Bleichrodt and Filko (2008) tested intertemporal additivity in a design that controlled for violations of expected utility and found support for it at the aggregate level, but much less so at the individual level. It should also be emphasized that other studies found support for the existence of one unifying concept of utility, both within the health domain (Stalmeier and Bezembinder, 1999) and outside the health domain (Abdellaoui et al., 2007a). All in all, the question of the transferability of utility is still open.

Our study is based on two implicit assumptions. First, that health policy decisions should as far as possible reflect the preferences of those who will be affected by them. Second, that ranking and choice reflect individual preferences. While the former assumption seems plausible, the latter is open to criticism. Ranking and choosing between health alternatives are not easy tasks and subjects are likely to have made errors and have adopted heuristics to facilitate these tasks. The previously observed discrepancies between ranking and choice suggest that ranking and choice induce different heuristics and biases (Bateman et al., 2007). We do not claim that ranking and choice are the gold standards against which health utility measurements should be judged. We believe, however, that they both are to some degree reflective of peoples' preferences. The fact that our findings are the same for ranking and for choice, in spite of the fact that these procedures tend to be affected by different cognitive biases, lends credibility to the robustness of our findings.

6. Conclusion

Prospect theory leads to better health evaluations than expected utility for decisions under risk. This finding does not necessarily translate to decision contexts other than risk. Our findings caution against the common assumption that health utilities can be used in different decision contexts. The SG has lower external validity either than the TTO (for intertemporal decisions) or than the SG adjusted for prospect theory (for decisions under risk). Our data add to the evidence indicating that the SG and algorithms based on it such as the SF-6D and the HUI should better be avoided when valuing health.

Acknowledgements

Jason Doctor and two reviewers gave many helpful comments. Jose Maria Abellan-Perpiñan and Jose Luis Pinto-Prades acknowledge financial support from Ministerio de Educación grant

SEJ2007-67734/ECON and from Consejería de Innovación (Junta de Andalucía) grant PO7-SEJ-02936. Han Bleichrodt's research was made possible through a VICI research grant from the Netherlands Organization for Scientific Research.

References

Abdellaoui, M., 2000. Parameter-free elicitation of utility and probability weighting functions. Management Science 46, 1497–1512.

Abdellaoui, M., Barrios, C., Wakker, P.P., 2007a. Reconciling introspective utility with revealed preference: experimental arguments based on prospect theory. Journal of Econometrics 138, 356–378.

Abdellaoui, M., Bleichrodt, H., Paraschiv, C., 2007b. Measuring loss aversion under prospect theory: a parameter-free approach. Management Science 53, 1659–1674.

Arrow, K.J., 1951. Alternative approaches to the theory of choice in risk-taking situations. Econometrica 19, 404–437.

Bateman, I., Day, B., Loomes, G., Sugden, R., 2007. Can ranking techniques elicit robust values? Journal of Risk and Uncertainty 34, 49–66.

Bleichrodt, H., 2002. A new explanation for the difference between standard gamble and time trade-off utilities. Health Economics 11, 447–456.

Bleichrodt, H., Abellan-Perpiñan, J.M., Pinto-Prades, J.L., Mendez-Martinez, I., 2007. Resolving inconsistencies in utility measurement under risk: tests of generalizations of expected utility. Management Science 53, 469–482.

Bleichrodt, H., Filko, M., 2008. New tests of QALYs when health varies over time. Journal of Health Economics 27, 1237–1249.

Bleichrodt, H., Johannesson, M., 1997. Standard gamble, time trade-off, and rating scale: experimental results on the ranking properties of QALYs. Journal of Health Economics 16, 155–175.

Bleichrodt, H., Pinto, J.L., 2000. A parameter-free elicitation of the probability weighting function in medical decision analysis. Management Science 46, 1485–1496.

Bleichrodt, H., Pinto, J.L., 2005. The validity of QALYs under non-expected utility. Economic Journal 115, 533–550.

Bleichrodt, H., Pinto, J.L., Wakker, P.P., 2001. Making descriptive use of prospect theory to improve the prescriptive use of expected utility. Management Science 47, 1498–1514.

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

Dyer, J.S., Sarin, R.K., 1982. Relative risk aversion. Management Science 28, 875–886.

Fishburn, P.C., 1989. Retrospective on the utility theory of von Neumann and Morgenstern. Journal of Risk and Uncertainty 2, 127–158.

Gafni, A., Birch, S., Mehrez, A., 1993. Economics, health, and health economics: HYEs versus QALYs. Journal of Health Economics 12, 325–339.

Gonzalez, R., Wu, G., 1999. On the form of the probability weighting function. Cognitive Psychology 38, 129–166.

Harsanyi, J.C., 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. Journal of Political Economy 63, 309–321.

Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47, 263–291.

Köbberling, V., Wakker, P.P., 2005. An index of loss aversion. Journal of Economic Theory 122, 119–131.

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.

Miyamoto, J.M., Wakker, P.P., Bleichrodt, H., Peters, H.J.M., 1998. The zero-condition: A simplifying assumption in QALY measurement and multiattribute utility. Management Science 44, 839–849.

Morrison, G.C., 2000. The endowment effect and expected utility. Scottish Journal of Political Economy 47, 183–197.

Richardson, J., 1994. Cost utility analysis: What should be measured? Social Science and Medicine 39, 7–22.

Robinson, A., Loomes, G., Jones-Lee, M., 2001. Visual analog scales, standard gambles, and relative risk aversion. Medical Decision Making 21, 17–27.

Rutten-van Mölken, M.P., Bakker, C.H., van Doorslaer, E.K.A., van der Linden, S., 1995. Methodological issues of patient utility measurement. Experience from two clinical trials. Medical Care 33, 922–937.

Stalmeier, P.F.M., Bezembinder, T.G.G., 1999. The discrepancy between risky and riskless utilities: a matter of framing? Medical Decision Making 19, 435–447.

Stalmeier, P.F.M., Wakker, P.P., Bezembinder, T.G.G., 1997. Preference reversals: violations of unidimensional procedure invariance. Journal of Experimental Psychology: Human Perception and Performance 23, 1196–1205.

Starmer, C., 2000. Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. Journal of Economic Literature 28, 332–382.

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 and comparison with standard gamble scores. Medical Decision Making 14, 82–90.

Sutherland, H.U., Llewellyn-Thomas, H., Boyd, N.F., Till, J.E., 1982. Attitudes toward quality of survival: the concept of 'maximal endurable time'. Medical Decision Making 2, 299–309.

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

van der Pol, M.M., Cairns, J., 2002. A comparison of the discounted utility model and hyperbolic discounting models in the case of social and private intertemporal preferences for health. Journal of Economic Behavior and Organization 49, 79–96.

van Osch, S.M.C., van den Hout, W.B., Stiggelbout, A.M., 2006. Exploring the reference point in prospect theory: gambles for length of life. Medical Decision Making 26, 338–346.

van Osch, S.M.C., Wakker, P.P., van den Hout, W.B., Stiggelbout, A.M., 2004. Correcting biases in standard gamble and time tradeoff utilities. Medical Decision Making 24, 511–517.

Wakker, P.P., 1994. Separating marginal utility and probabilistic risk aversion. Theory and Decision 36, 1–44.

Weber, M., Eisenführ, F., von Winterfeldt, D., 1988. The effects of splitting attributes on weights in multiattribute utility measurement. Management Science 34, 431–445.