DOCUMENTO DE TRABAJO. E2007/ 14

Testing the predective validity of the time trade-off and the Standard Gamble

JOSE MARÍA ABELLÁN-PERPIÑAN
HAN BLEICHRODT
JOSÉ LUIS PINTO PRADES

Centro de Estudios Andaluces
CONSEJERÍA DE LA PRESIDENCIA

El Centro de Estudios Andaluces es
una entidad de carácter científico y
cultural, sin ánimo de lucro, adscrita
a la Consejería de la Presidencia
de la Junta de Andalucía.
El objetivo esencial de esta institución
es fomentar cuantitativa y cualitativamente
una línea de estudios e investigaciones
científicas que contribuyan a un más
preciso y detallado conocimiento de
Andalucía, y difundir sus resultados
a través de varias líneas estratégicas.

El Centro de Estudios Andaluces desea
generar un marco estable de relaciones
con la comunidad científica e intelectual
y con movimientos culturales en
Andalucía desde el que crear verdaderos
canales de comunicación para dar
cobertura a las inquietudes intelectuales y culturales.

Las opiniones publicadas por los autores en
esta colección son de su exclusiva responsabilidad

Depósito Legal: SE- 6753-07

JUNTA DE ANDALUCIA
Centro de Estudios Andaluces
CONSEJERÍA DE LA PRESIDENCIA

ECO2007/14

Testing the Predictive Validity of the Time Trade-Off and the Standard Gamble

José María Abellán-Perpiñan
Universidad de Murcia

Han Bleichrodt
Erasmus University Rotterdam

José Luis Pinto-Prades *
Universidad Pablo de Olavide
Centro de Estudios Andaluces

RESUMEN

Este trabajo estudia la consistencia de los métodos de obtención de preferencias con las preferencias individuales. Comparamos tres métodos, la Compensación Temporal, la Lotería Estandar y una versión de la Lotería Estandar que corrige las desviaciones de la utilidad esperada modelizada por prospect theory. Las preferencias individuales se miden tanto a través de ordenaciones como a través de elecciones. En las decisiones que no implican riesgo la Compensación Temporal es más consistente con las preferencias de la gente, ocupando la Lotería Estándar la segunda posición. En las decisiones en las que hay riesgo, la Lotería Estándar corregida es más consistente con las preferencias individuales. Nuestros datos no corroboran el habitual supuesto de la economía de la salud de que la utilidad es transferible entre contextos de decisión.

Health utility measurement, QALYs, standard gamble, time trade-off, prospect theory.

Palabras clave: medida de la utilidad en salud, lotería estándar, compensación temporal, prospect theory

[†] Autor para correspondencia: jluis.pinto.ext@centrodeestudiosandaluces.es

ABSTRACT

This paper tests the consistency of health utility measurements with individual preferences. We compare three methods, the time trade-off, the standard gamble and a version of the standard gamble that corrects for the deviations from expected utility modelled by prospect theory. Individual preferences are measured both through a ranking task and through a choice task. In decisions involving no risk the time trade-off is most consistent with people's preferences with the standard gamble a close second. In decisions involving risk the corrected standard gamble is most consistent with people's preferences. Our data do not support the common assumption in health economics that utility is transferable across decision contexts.

Keywords: Health utility measurement, QALYs, standard gamble, time trade-off, prospect theory.

JEL Classification: I10

1. Introduction

A well-known problem in health utility measurement is that the main elicitation methods, the standard gamble (SG) and the time trade-off (TTO)[1], lead to systematically different results. The common pattern is that the SG exceeds the TTO (Dolan, 2000). A danger of this divergence in elicited utilities is that the outcomes of economic evaluations of health care may come to depend on the elicitation method used. This would not be a reason for concern if it were known which method should be preferred. Unfortunately, there exists no consensus on this question.

The traditional belief was that the SG should be preferred, because it is based on expected utility, the dominant prescriptive theory of decision under risk. Economic evaluations of health care are a prescriptive exercise, the purpose being to prescribe how health care allocation decisions should be made, and, hence health utility measurements should be based on the prevailing prescriptive theory of decision making, expected utility. The problem with this argument is that health utility measurement is a descriptive task and the descriptive deficiencies of expected utility are widely documented (Starmer, 2000). Evidence of violations of expected utility for health outcomes was obtained by Llewellyn-Thomas et al. (1982), Rutten-van Mölken et al. (1995), and Oliver (2003) amongst others. If the SG is used in spite of the known deviations from expected utility, health utility measurements are likely to be biased. Evidence of such biases is reported in Bleichrodt et al. (2007), who showed that the SG leads to utilities that are too high.

Bleichrodt and Johannesson (1997) tested to what extent SG and TTO were consistent with people's preferences over health profiles and found that the TTO most closely mirrored these. An explanation for the higher consistency of the TTO

[1] Other commonly used methods are the visual analogue scale and the person trade-off. We do not consider these methods in this paper.

with people's preferences was offered by Bleichrodt (2002). That paper argued that SG and the TTO utilities are affected by four factors, utility curvature and three deviations from expected utility: probability weighting, loss aversion, and scale compatibility. These factors lead to an upward bias in the SG, but have offsetting effects on TTO utilities. Bleichrodt's (2002) analysis implies that discounting will reduce the consistency of the TTO with people's preferences, a conclusion that born out by the data in Bleichrodt and Johannesson (1997).

The purpose of this paper is to perform a new test of the consistency of SG and TTO with individual preferences. There are at least three motivations for such a new test. A first motivation is that the study by Bleichrodt and Johannesson (1997) was somewhat limited in scope. They considered only one health state and only health profiles that involved no risk. Because the TTO involves no risk either, the evaluation of the riskless profiles may have been cognitively similar to the evaluation of the TTO questions enhancing the consistency of the TTO with preferences. In the present study we consider different health states and both profiles involving risk and profiles involving no risk.

Second, Bleichrodt and Johannesson (1997) measured people's preferences by asking them to rank the health profiles and took this ranking as the "gold standard" against which the performance of SG and TTO was judged. Ranking health profiles reflects the purpose of economic evaluation which is to establish priorities in health care through QALY-league tables. However, a case could also be made for using choice instead of ranking as choice is the basic primitive of economics. Recent research has shown that ranking and choice can produce substantially different results (Bateman et al., 2006, Oliver, 2006). In this paper we will elicit people's preferences both through ranking and through choice and we will examine to what extent they lead to different conclusions about the relative performance of the SG and the TTO.

A final motivation for this study is that since Bleichrodt and Johannesson (1997) improvements of the SG have been developed. Bleichrodt, Pinto, and Wakker (2001) proposed formulas that correct for the biases in the SG caused by probability weighting and loss aversion and showed that these corrections were able to remove inconsistencies in SG measurements. Further support for these corrections is in van Osch et al. (2004), van Osch, van den Hout, and Stiggelbout (2006), and Bleichrodt et al. (2007). The findings of van Osch et al. (2004) suggest that the TTO also leads to utilities that are too high and that the corrected SG might better represent people's preferences for health. Hence, it of interest to explore whether we can improve the consistency of economic evaluations of health care with individual preferences by using the corrected SG instead of the TTO or the uncorrected SG.

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 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 (p:q; q′) 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 $(p{:}q;\, q')$ that q is at least as good as q′: $q \succcurlyeq q'$.

Expected utility holds if prospects $(p{:}q;\, q')$ are evaluated as $pU(q) + (1-p)U(q')$ and preferences and choices correspond with this evaluation. U is a utility function over chronic health states that is unique up to unit and location, i.e., we can freely select the utility of two health states.

Prospect theory (Tversky and Kahneman, 1992), currently the main descriptive theory of decision under risk, 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 from a reference point. People are assumed to be more sensitive to losses than to commensurate gains, a phenomenon known as *loss aversion*. Sign-dependence also affects the weighting of probabilities: prospect theory allows that probability weighting for gains is different from probability weighting for losses.

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 (p:(FH,T); Death), where FH denotes full health. Under expected utility, this indifference implies that

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

Under prospect theory, the evaluation of a binary prospect depends on the sign of its outcomes. Several studies have provided evidence that people evaluate SG questions by taking the sure outcome (Q,T) as their reference point (Morrison, 2000, Bleichrodt, Pinto, and Wakker, 2001, Robinson, Loomes, and Jones-Lee, 2001, van Osch et al., 2004). This implies that the risky prospect (p:(FH,T); Death) is *mixed*, i.e., it yields a gain (FH,T) with probability p and a loss Death with probability 1−p. The evaluation of the SG question is then equal to

$$w^{+}(p)\big(U(FH,T) - U(Q,T)\big) - \lambda w^{-}(1-p)\big(U(Q,T) - U(Death)\big) = 0, \qquad (2)$$

where w^+ denotes the probability weighting function for gains, w^- denotes the probability weighting function for losses, λ is a coefficient that reflects loss aversion and U is a utility function over chronic health states that is unique up to unit and location.

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

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

where λ_t is a decision 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 *nonlinear QALY model.* Special cases of Eq.(3) are the *linear QALY model*, for which $\lambda_t = 1$ for all t, and the QALY model with constant discounting, for which $\lambda_t = 1/(1+r)^{t-1}$ for all t. Both special cases are widely used in applied economic evaluations of health care.

Define $L(s) = \sum_{t=1}^{s}\lambda_t$. L can be interpreted as a utility function over life duration. Equation 3 implies that U in Eqs. 1 and 2 is equal to U(Q,T) = H(Q)*L(T). Preference conditions for the nonlinear 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 nonlinear QALY model implies that more years in full health are preferred to less, something that we will use in what follows.

Under the nonlinear QALY model and the common scaling H(FH) = 1 and U(Death) = 0, Eq.(1) yields H(Q) = p, 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) , we obtain that

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

To compute H(Q) we must know the probability weighting functions and the loss aversion coefficient. Here we will assume the estimates obtained by Tversky and Kahneman (1992). Bleichrodt et al. (2001) used these estimates to measure H(Q) and found that at the aggregate level they performed well. 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).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 *corrected SG utilities.*

The *TTO* elicits the number of years T_1 in full health that makes a decision maker indifferent to T years in some impaired health state Q. Under the QALY model and the scaling H(FH) = 1 this implies that $H(Q) = L(T_1)/L(T)$. In the empirical literature on the TTO L(·) is generally assumed linear. Then $H(Q) = T_1/T$.

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 there is no unifying concept of utility and utility is context-specific: The utility function that is elicited under risk may be different from the function that is elicited under certainty (Arrow, 1951, Dyer and Sarin, 1982, Fishburn, 1989, Gafni, Birch, and Mehrez, 1993). Others have argued in favor of the existence of one unifying concept of utility (Harsanyi, 1955, Richardson, 1994, Wakker, 1994). For empirical evidence that utility in different decision contexts is similar see, for example, Abdellaoui, Barrios, and Wakker (2007). In health economics it is very common to use the same utilities in different contexts. For example, SG utilities are measured under risk but are applied in societal decisions about the allocation of health care resources, i.e. in welfare evaluations. As another example, time trade-off utilities are measured in an intertemporal setting but are used both in decisions under risk and in welfare evaluations. Our empirical results will shed additional light on the question to what extent this common practice is justified.

3. Experiment

General idea–– We computed the number of QALYs of 10 health profiles, 5 riskless and 5 risky, based on TTO, SG, and corrected SG and compared the implied ranking of the health profiles with the directly elicited ranking and with the ranking implied by directly observed choices.

Subjects–– The subjects were sixty-five economics students (aged between 22 and 29) from the University of Murcia. They were paid ú36 to particip

experimental sessions. In this paper we will 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 so that recall bias is unlikely. Prior to the actual experiment, the questionnaire was tested in several pilot sessions.

Stimuli––We elicited the utility of the EQ-5D health states 22122 and 22322. The description of the health states is given in Table 1. Throughout the experiment, the health states were labeled A and B. Health state A dominates health state B in the sense that it yields on each dimension an outcome that is at least as good as the corresponding outcome of B. Full health was described as no limitations on any of the dimensions.

Table 1: The description of health states A and B

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

We asked six questions for each of the SG and the TTO[2] by combining the two health states with three different values for T: 13 years, 24 years, and 38 years. We used durations less 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.

[2] Recall that the corrected SG can be computed from the response to the SG question.

Table 2: The health profiles used

Riskless profiles
14A+3FH 9A+4B+4FH 4FH+13B 1FH+13B+3FH 2FH+4A+8B
Risky profiles
(0.63:11A; Death) (0.5:17A; Death) (0.5:11B; 7A) (0.5:6B; 6FH) (0.45: 14A; 7B)

Table 2 displays the health profiles used in the experiment. We used five riskless and five risky health profiles. The notation 9A+4B+4FH stands for nine year in health state A followed by 4 years in health state B followed by 4 years in full health. We only used health profiles of constant quality in the risky profiles to keep the tasks tractable. All health profiles ended with death. The health profiles were printed on cards. The riskless health profiles were displayed as stacked bars with the size of each component of the bar corresponding to the duration of the health state. Figure 1 gives an example (translated from Spanish into English). The risky health profiles were displayed as pie charts with the size of each pie corresponding to the size of the probability. Figure 2 gives an example.

Figure 1: Display of the riskless health profiles

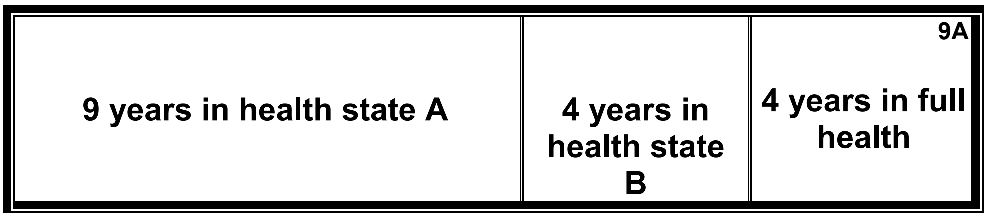
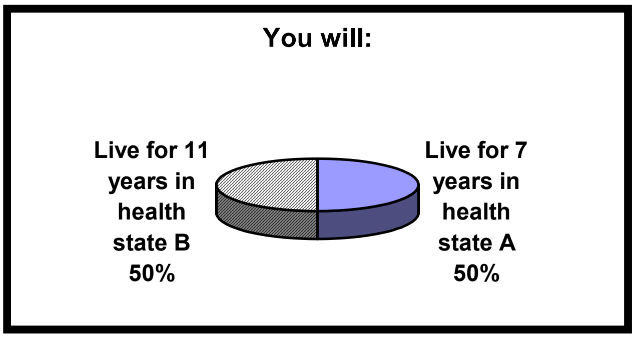


Figure 2: Display of the risky health profiles



Procedures–– Preferences were elicited through a ping-pong 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 "zoomed in" on the parameter value for which subjects were indifferent between the alternatives.

Recruitment of subjects took place one week before the actual experiment started. At recruitment, subjects received information about the experiment and were asked to read the descriptions of the two health states. In addition, the subjects were handed a practice question on the SG method. They were asked to answer this practice question at home. This procedure was intended to familiarize them 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 he understood the task. 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 the question before the actual experiment started.

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 section. To minimize response errors, the subjects had to confirm the elicited indifference value after each 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 profiles 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 profile they preferred. To reduce the cognitive burden, we only asked subjects to compare riskless with riskless profiles and risky with risky profiles. Hence, they did not compare riskless profiles with risky profiles.

Analysis–– Corrected SG utilities were computed from the response 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 (uncorrected SG, corrected 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 profiles.

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

$$L(9)H_{SG,24}(A) + \big(L(13)–L(9)\big)H_{SG,24}(B) + \big(L(17)–L(13)\big), \qquad (5)$$

where $H_{SG,24}$ denotes the health state utility according to the uncorrected SG determined in the question with T = 24 years. The number of QALYs of the risky health profiles was computed as the expected utility of the risky health profile. So for example, the number of QALYs of the risky health profile (0.45: 14A; 7B) according to the corrected SG and T = 38 yearswas computed as

$$0.45*L(14)*H_{CSG,38}(A) + 0.55*L(7)*H_{CSG,38}(B). \qquad (6)$$

We only applied prospect theory to correct the health state utilities and we did not apply prospect theory in the evaluation of the risky profiles for the following reason. The measurement of health state utilities is a descriptive task and it is here that we should correct for the deviations from expected utility. However, once the health state utilities and the probabilities of the different states of nature are known, the final evaluation of health profiles is a normative exercise and should be based on a normative theory. Expected utility is still unchallenged as a normative theory and, hence, we used expected utility to calculate the expected number of QALYs.

To compute Eqs. (5) and (6) we must know L. We first assumed that L was linear, i.e., we assumed the linear QALY model. The reason to perform this analysis was the widespread use of the QALY model. The (expected) number of QALYs according to the three methods implied a ranking of the riskless and the risky health profiles. This ranking could be compared with the directly elicited ranking. The consistency of the three methods with the directly elicited ranking was assessed both by the Spearman rank correlation coefficient and by Kendall's tau. The elicited pairwise choices also led to a rank ordering of the riskless health profiles and a rank ordering of the risky health profiles in the sense that the rank of a health profile was determined by the number of times it was chosen over another health profile. A problem in determining the rank ordering of the health states by the choice-based procedure is that it may not be unique because intransitive choice patterns were possible. For example, a subject could prefer 9A+4B+4FH to 1FH+13B+3FH, 1FH+13B+3FH to 2FH+4A+8B, but 2FH+4A+8B to 9A+4B+4FH. The relative ranking of these profiles is then not clear. Subjects who exhibited intransitive choices were excluded from the analysis of the choice data.

To exclude the possibility that our results were driven by violations of the QALY model, we also analyzed the data allowing for utility curvature. To model utility curvature, 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 test for robustness, allowing for the fact that people may not behave according to constant discounting either (van der Pol and Cairns, 2002), we also evaluated the data under a power specification for L. The results were similar, but convergence was better for the exponential specification and, hence, we will report those in Section 4.

All statistical tests are parametric tests. We also performed nonparametric tests but these led to the same results and, hence, they are not reported separately.

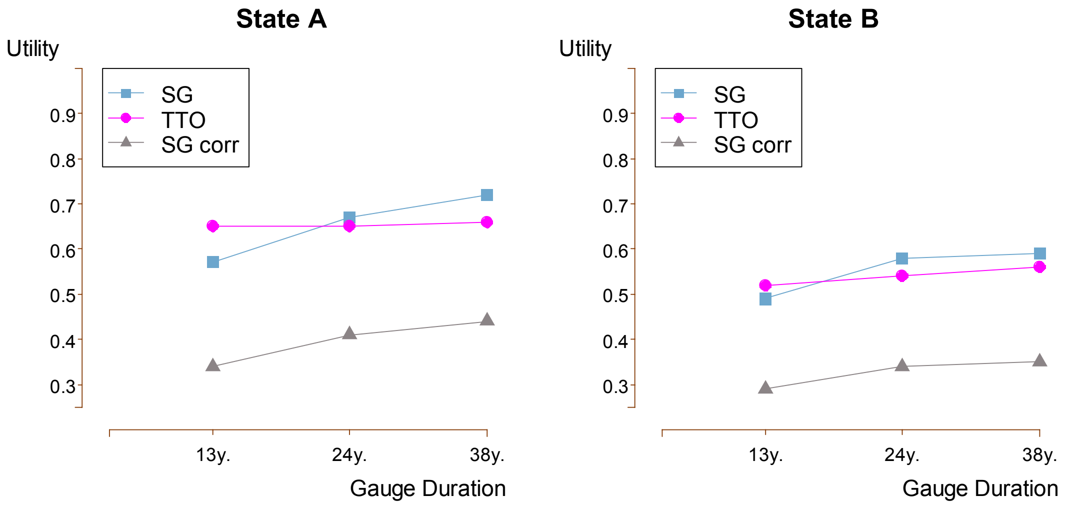
4. Results

Nineteen of our 65 subjects were excluded because their responses implied that they did not always prefer more life-years to less. As expected, this happened primarily for the less attractive health state B. Only two subjects violated monotonicity with respect to life duration for health state A. The seventeen subjects who violated monotonicity with respect to life duration only for health state B had similar SG and corrected SG valuations for health state A as the subjects in our analysis. Their TTO valuations were, however significantly lower. Hence, 46 subjects subjects were included in the analysis of the ranking data.

We had to exclude 5 more subjects in the analysis of the choices between the riskless profiles, because they violated transitivity. Hence, 41 subjects were included in the analysis of the riskless choices. In the analysis of the choices between the risky profiles we had to exclude 11 subjects because of intransitive choices. Hence, 35 subjects were included in the analysis of the risky choices. No subject had to be excluded both in the analysis of the choices between the riskless profiles and in the analysis of the choices between the risky profiles. This suggests that the observed intransitivities are caused by errors and are not inherent characteristics of subjects' preferences. That more subjects were excluded in the analysis of the risky choices may reflect that subjects perceived these choices as more complicated.

Figure 3 shows the mean utilities of health states A and B according to the three methods used. The medians were similar. The utility of health state A was higher than that of health state B according to all three methods as should be 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 corrected SG, the difference being around 0.25. The difference between SG and TTO with corrected SG is significant ($p < 0.001$). SG and TTO differ significantly only 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 ($p = 0.023$).

Figure 3: Mean utilities for health states A and B according to the three methods



Riskless profiles—— Table 3 shows the mean rank of the five riskless profiles (column Actual) both based on the ranking data and based on the choice data. Higher ranks reflect more attractive profiles. The aggregate data on rankings and choices are to a large extent comparable except that 14A+3FH appears more attractive and 4FH+13B less attractive in the choice data. The mean rank of the profile 4FH+13B was 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 was even ranked higher than 4FH+13B suggesting negative discounting, i.e., convex utility for life duration.

Table 3 : Mean ranks of the riskless profiles according to the three methods

	Ranking				Choice			
Profile	**Actual**	**TTO**	**SG**	**Corr. SG**	**Actual**	**TTO**	**SG**	**Corr. SG**
14A+3FH	1.35	1.80	1.93	2.54	1.05	1.81	1.79	2.46
9A+4B+4FH	2.04	2.00	2.01	1.65	1.98	1.97	1.96	1.56
2FH+8A+4B	3.57	4.99	4.96	4.96	3.63	4.99	4.96	4.98
4FH+13B	4.00	2.58	2.51	2.41	4.46	2.58	2.59	2.47
1FH+13B+3FH	4.04	2.58	2.51	2.41	3.88	2.58	2.59	2.47

Figure 4 displays the mean Spearman rank correlation coefficients for the riskless profiles when we assume that the linear QALY model holds. The results based on Kendall's tau were similar. Figure 4A shows the results for the ranking data and figure 4B for the choice data.

Figure 4: Mean Spearman Rank Correlations for the Riskless Profiles

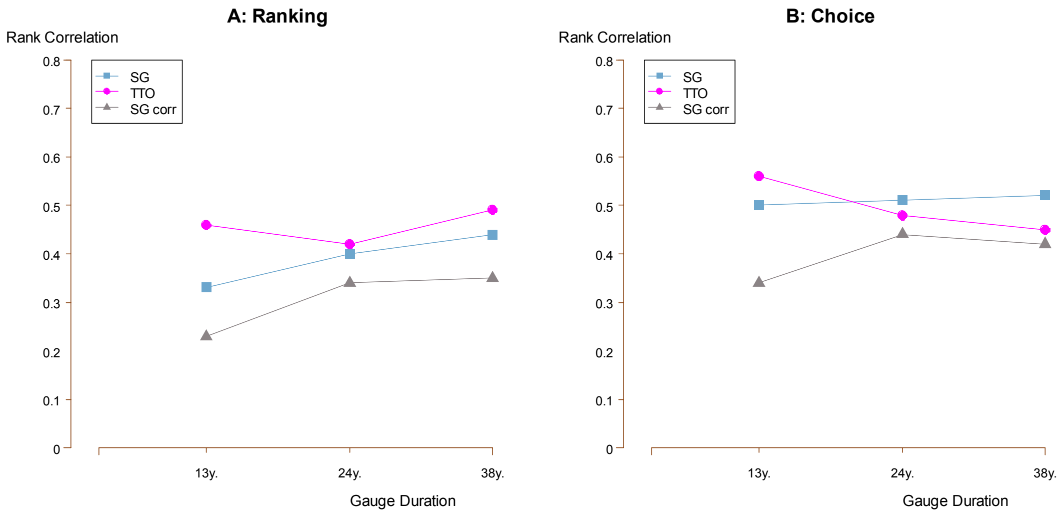


Figure 4A shows that QALYs based on the TTO were most consistent with the ranking data. This finding is in line with the main conclusion in Bleichrodt and Johannesson (1997). The difference between TTO and SG is, however, less pronounced than in Bleichrodt and Johannesson (1997) and does not reach statistical significance (p > 0.10 in all three comparisons). TTO and SG are more consistent with the ranking data than the corrected SG. The differences are, however, only significant for the TTO.

The choice data paint a somewhat different picture. The TTO is no longer more consistent with the choice data than the SG. In fact, the SG is more consistent than the TTO for durations 24 and 38 years. Moreover, even though TTO and SG are more consistent with the choice data than the corrected SG, the difference between TTO and corrected SG is small (and insignificant) for durations 24 years and 38 years.

Table 3 shows the predicted ranks of the profiles according to the three methods. The results differ slightly for the ranking and the choice data because the number of subjects included differs between these data. One explanation for the data in the Table is that the three methods, and in particular the corrected SG, underestimate the utility of A and overestimate the utility of B. This can explain why the attractiveness of the profiles 4FH+13B and 1FH+13B+3FH is overestimated and that of the profile 14A+3FH is underestimated. The underestimation of the attractiveness of profile 2FH+8A+4B can also be explained by underestimation of the difference in utility between health states A and B.

Risky prospects—— Table 4 shows the mean ranks of the five risky prospects (column Actual) for the ranking and the choice data. The ranking and choice data are again comparable, except that prospect (0.5: 6B; 6FH) is more attractive in the choice data and (0.5: 11B; 7A) in the ranking data. The profiles involving the possibility of death are considered the least attractive. The profile (0.5:6B; 6FH) 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. The prospect (0.5: 17 A; Death) was considered unattractive suggesting that subjects did not value additional life duration as much as the linear QALY model assumes. Note that this finding is not caused by considerations of maximal endurable time as the subjects who violated monotonicity with respect to life duration were excluded from the analyses.

Table 4: Mean ranks of the risky profiles according to the three methods

	Ranking				**Choice**			
Profile	**Actual**	**TTO**	**SG**	**Corr. SG**	**Actual**	**TTO**	**SG**	**Corr. SG**
(0.63:11A; Death)	3.78	4.36	4.37	4.88	3.71	4.43	4.45	4.90
(0.5:17A; Death)	4.61	2.35	2.40	3.09	4.43	2.42	2.47	3.16
(0.5:11B; 7A)	2.61	2.98	2.78	3.76	3.06	2.92	2.71	3.70
(0.5:6B; 6FH)	2.52	4.23	4.33	1.52	2.11	4.17	4.25	1.48
(0.45: 14A; 7B)	1.48	1.09	1.13	1.75	1.69	1.07	1.12	1.76

Figure 5 shows the mean Spearman rank correlation coefficients for the three methods. Part A displays the results for the ranking data, part B for the choice data. The corrected SG is substantially and significantly more consistent with subjects' directly elicited preferences than the TTO and the SG ($p < 0.001$). This holds both for the ranking and for the choice data, but the difference is particularly large for the choice data. No significant differences were observed between SG and TTO.

Figure 5: Mean Spearman Rank Correlations for the Risky Profiles

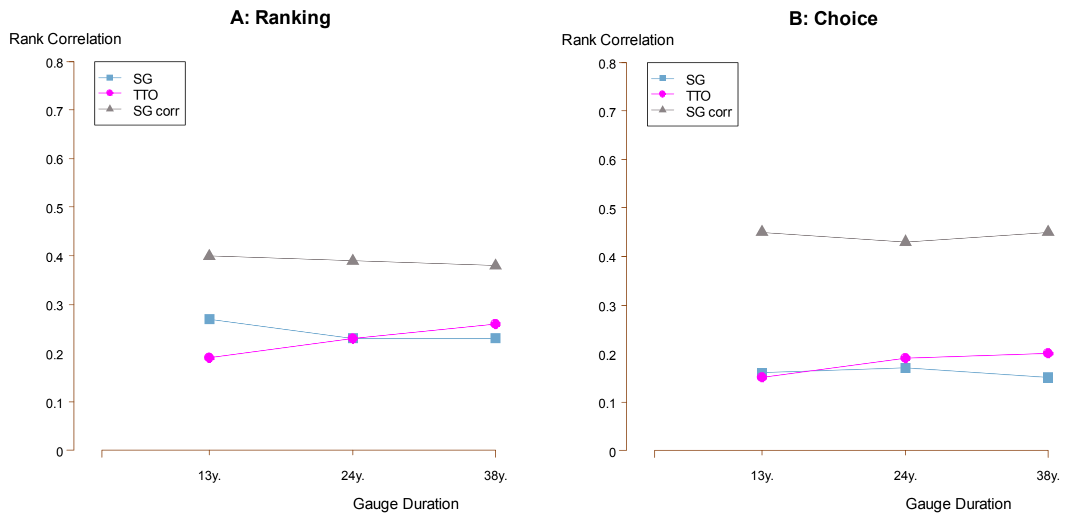
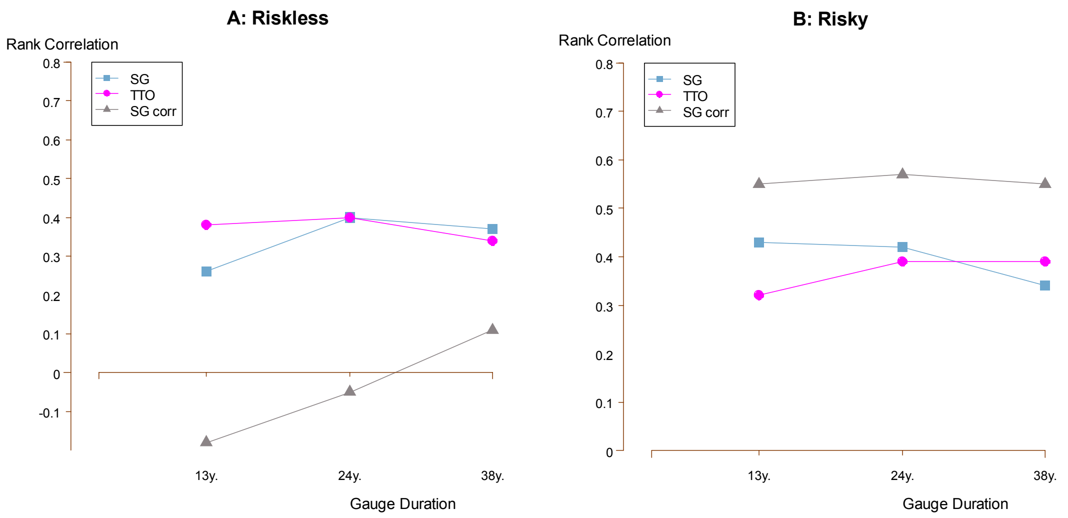


Table 4 also shows the predicted rankings according to the three methods. Again, the difference between predictions for the ranking and for the choice data is caused by the difference in the number of included subjects. The table shows that, compared with the actual ranking and the actual choices, all methods imply an overestimation of the attractiveness of the prospect (0.5: 17A; Death). This could be explained by an overestimation of the utility of living longer by the linear QALY model. This overestimation could also explain why the TTO and SG value (0.5: 6B; 6FH) much too low. Living fewer life-years is not as bad as the assumption of linear utility for life duration implies. That the corrected SG does not suffer from this problem might be caused by its underestimation of the difference in utility between A and B. A similar observation regarding the difference in utility between health states A and B was made in the riskless case.

Nonlinear utility—– 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, Cher, Miyamoto, and Lenert, 1997) and our data on the risky prospects also suggested 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 did not affect the conclusions. Figure 6 shows the results for the choice data when 5% discounting is applied. The results for the ranking data and for 3% discounting are similar. The patterns are comparable to those in Figures 4 and 5, which display the situation of no discounting. In general, discounting tends to reduce the rank correlation coefficients for the riskless data and increases the rank correlation coefficients for the risky data. This observation is confirmed when we estimated optimal discounting parameters for each subject. In general, for the riskless data no discounting was optimal. For the risky data the optimal parameters indicate strong discounting (> 30%).

Figure 6: Mean Spearman Rank Correlation Coefficients for the Choice Data with 5% Discounting



5. Discussion

A fundamental premise of welfare economics is that public policy decisions should as far as possible reflect the preferences of those who will be affected by them. This paper has explored the extent to which the common measures of health utility reflect people's preferences for health. The results suggest that the answer to this question depends on the decision context. For decisions involving no risk, the TTO performed best. The SG performed slightly worse for the ranking data, but performed equally well for the choice data. The corrected SG performed worse. For decisions under risk, however, the corrected SG was clearly more consistent with the data than SG and TTO. Given that most medical decisions involve risk, our findings suggest that in most medical decisions using the corrected SG will lead to recommendations that are more consistent with people's preferences than using SG or TTO. This is in line with the observation of van Osch et al. (2004) that TTO and SG give utilities that are too high in a risky decision task.

A difficulty of using the corrected SG is that it requires measurements of or assumptions about probability weighting and loss aversion. Previous studies have shown that at the aggregate level the parameter estimates of Tversky and Kahneman perform well. Our study adds to this evidence by showing that the correction of the SG for probability weighting and loss aversion led to a significant improvement in consistency with subjects' preferences as compared with the uncorrected SG and the TTO. Hence, at the aggregate level the corrected SG is more consistent with subjects' preferences and is equally tractable as the uncorrected SG and the TTO. Previous studies have shown, however, that at the individual level much variation exists in probability weighting and loss aversion. Consequently, in medical decision making where individual treatment recommendations have to be made, the parameters of

Tversky and Kahneman (1992) may not always work well and individual measurements are needed. There exist methods to measure probability weighting and loss aversion (Abdellaoui, 2000, Abdellaoui, Bleichrodt, and Paraschiv, 2007, Bleichrodt and Pinto, 2000) but these require asking extra questions and, hence, increase the cognitive burden for subjects.

Our data suggest that utility is context-specific and that there is no unifying concept of utility that can explain our data. In the context of certainty, TTO and uncorrected SG were more consistent with subjects' directly elicited preferences, whereas in the context of risk the corrected SG was more consistent. Utilities under certainty are higher than utilities under risk and our data suggest that one should be careful in transferring utility estimates across decision contexts. This is an important finding for economic evaluations of health care where such transferability is commonly assumed. It should be mentioned though that other studies found support for the existence of one unifying concept of utility (Abdellaoui, Barrios, and Wakker, 2007). All in all, the question of the existence of one unifying concept of utility seems still open. More research on this is needed.

An implicit assumption of our study is that ranking and choice reflect individual preferences. This assumption is probably too strong. Ranking and choosing between health profiles are not easy tasks and subjects are likely to 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., 2006, Oliver 2006). 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.

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., 2007. Reconciling introspective utility with revealed preference: Experimental arguments based on prospect theory. Journal of Econometrics 138, 356-378.

Abdellaoui, M., Bleichrodt H., Paraschiv C., 2007. 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. C., Sugden R., 2006. Ranking versus choice in the elicitation of preferences. Working Paper, University of East Anglia.

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 J. M., Pinto J. L., Mendez I., 2007. Resolving inconsistencies in utility measurement under risk: Tests of generalizations of expected utility. Management Science 53, 469-482.

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., 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.

Cher, D. J., Miyamoto J., Lenert L. A., 1997. Incorporating risk attitude into Markov-process decision models. Medical Decision Making 17, 340-350.

Dolan, P., 2000. The measurement of health-related quality of life for use in resource allocation decisions in health care. In: Culyer, A. J. and Newhouse, J. P. (Eds.), Handbook of health economics, vol. 1b. Elsevier Science, Amsterdam, 1723-1760.

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.

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.

Oliver, A. J., 2003. The internal consistency of the standard gamble: Tests after adjusting for prospect theory. Journal of Health Economics 22, 659-674.

Oliver, A. J., 2006. Further evidence of preference reversals: Choice, valuation and ranking over distributions of life expectancy. Journal of Health Economics 25, 803-820.

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.

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.

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.