Journal of Health Economics 27 (2008) 1237–1249

Contents lists available at ScienceDirect

Journal of Health Economics

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

New tests of QALYs when health varies over time

Han Bleichrodt [a,*], Martin Filko [b]

[a] Department of Economics & iMTA/iBMG, Erasmus University, H13-27, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands
[b] Department of Health Policy and Management, Erasmus University, Rotterdam, The Netherlands

ARTICLE INFO

Article history:
Received 5 December 2007
Received in revised form 21 March 2008
Accepted 15 May 2008
Available online 21 July 2008

JEL classification:
I10

Keywords:
QALYs
Economic evaluation of health care
Decision under risk
Nonexpected utility

ABSTRACT

This paper performs new tests of the QALY model when health varies over time. Our tests do not involve confounding assumptions and are robust to violations of expected utility. The results support the use of QALYs at the aggregate level, i.e. in economic evaluations of health care. At the individual level, there is less support for QALYs. The individual data are, however, largely consistent with a more general QALY-type model that remains tractable for applications.

1. Introduction

Quality-adjusted life-years (QALYs) are the most widely used measure of health in economic evaluations of health care. According to the QALY model, the utility of a health profile equals the sum of the utilities of its constituent health states. The popularity of QALYs can be explained by their tractability and intuitive appeal: QALYs are easy to use and easy to explain to policy makers. A drawback of QALYs may be that they are too simple and do not represent people's preferences for health in a reliable manner. An obvious danger of using an unreliable measure in economic evaluations of health care is that treatment recommendations and reimbursement decisions are made that do not represent people's interests.

Several studies have tested the validity of QALYs when health states are chronic (for an overview see Bleichrodt and Pinto-Prades, 2006). Less evidence exists on the validity of QALYs for the more realistic case where health varies over time. Most of the existing studies tested the validity of QALYs by comparing the directly elicited utility of a health profile with the indirect utility that is obtained by adding the utilities of the independently rated constituent health states. The evidence from these studies is mixed with some studies finding large and significant differences (e.g. Richardson et al., 1996) and others finding only small and typically insignificant differences (Mackeigan et al., 1999; Brazier et al., 2006). The performance of the QALY model is better at the aggregate level than at the individual level (e.g. Kuppermann et al., 1997) although Krabbe and Bonsel (1998) found that only a small proportion of their subjects violated additivity.

There are several problems with the above method for testing the validity of QALYs. A first problem is that confounding assumptions must be made, in particular about the discounting of future health. All the above studies assumed constant discounting. Empirical evidence abounds, however, that the descriptive record of constant discounting is poor and that

* Corresponding author.
E-mail addresses: bleichrodt@few.eur.nl (H. Bleichrodt), filko@bmg.eur.nl (M. Filko).

0167-6296/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhealeco.2008.05.009

people deviate from it systematically (Frederick et al., 2002; van der Pol and Cairns, 2002). The problem of confounding assumptions is that when a difference between the direct and the indirect valuation of a profile is observed we do not know what is causing this difference and, hence, no information is obtained on how QALYs might be improved.

A second problem is that the valuation of health profiles and the valuation of health states involve different experimental stimuli and may, therefore, invoke different cognitive processes. Consequently, they may be susceptible to different decision biases. In particular, a cognitively demanding task such as the valuation of health profiles may induce the use of simplifying heuristics.

A third problem arises if the measured health utilities are biased. Many of the abovementioned studies used the standard gamble. It is well known that the standard gamble leads to utilities that are too high (van Osch et al., 2004; Bleichrodt et al., 2007; Doctor et al., in press). This upward bias is only present once in the direct valuation of the health profiles, but affects the valuation of each of the constituent health states and, hence, it is present more than once in the indirect valuation of the health profiles based on the utilities of their constituent health states. Consequently, we would expect that the estimation of the utility of a health profile from its constituent health states exceeds the direct valuation of the profile when the standard gamble is used and this is indeed what is typically observed.

The above problems can be avoided by testing the preference conditions on which QALYs are based. This approach was adopted by Treadwell (1998), who tested preference independence, and Spencer and Robinson (2007), who tested utility independence. Preference independence and utility independence are implied by the QALY model, i.e. they are necessary conditions for the QALY model. Both Treadwell (1998) and Spencer and Robinson (2007) found that the condition they tested was generally supported. The support for these conditions does not imply, however, that the QALY model holds as the conditions are also consistent with other, more general, decision models. To obtain conclusive evidence on the validity of QALYs, conditions must be tested that are both implied by the QALY model and that imply the QALY model, i.e. conditions that are both necessary and sufficient.

Spencer (2003) tested such a condition. She observed a violation of this condition at the individual level, but the violation was not systematic and might just be due to noise. Spencer's test is only valid if people behave according to expected utility. It is well known, however, that people systematically deviate from expected utility (Starmer, 2000). Hence, it cannot be excluded that the violations of the QALY model that Spencer observed reflected violations of expected utility rather than violations of the QALY model. To wit, while many studies observed violations of the QALY model for chronic health states under expected utility, Doctor et al. (2004) found no violations of the QALY model when violations of expected utility were taken into account.

In this paper we provide new tests of the QALY model when health varies over time. Like Treadwell (1998), Spencer (2003), and Spencer and Robinson (2007) we test preference conditions and, hence, our tests are not affected by the problems surrounding the comparison between direct and indirect valuations of health profiles. We test two conditions. The first condition, generalized marginality, is the central condition underlying the QALY model and implies that health profiles can be evaluated additively. Hence, like Spencer (2003) our test is both necessary and sufficient for the QALY model. An important difference with Spencer (2003) is that our test does not assume expected utility but is valid under a more general utility model that includes many of the theories of decision under risk that exist today. Hence, our tests are robust to violations of expected utility.

As generalized marginality is a restrictive condition and we could well imagine people violating it, we also tested utility independence. Utility independence is less restrictive than generalized marginality and it does not imply the QALY model. As will be explained in Section 2, utility independence still implies a model that is tractable and that can be used in practical applications. Spencer and Robinson (2007) also tested utility independence. Our experimental protocol differed in several respects from the protocol used by Spencer and Robinson (2007) and, hence, our data on utility independence complement Spencer and Robinson's analysis. Taken together the two studies provide insight into the validity of utility independence, an important condition for preference modeling and utility measurement.

The paper is structured as follows. Section 2 provides theoretical background and explains generalized marginality and utility independence. Section 3 describes the design of an experiment that aimed to test these conditions and Section 4 its results. Section 5 discusses the results and concludes.

2. Background

Let $q=(q_1,\ldots,q_T)$ denote a *health profile* where q_t stands for the health state at period t and T denotes the number of periods of survival. We assume that all health states are better than death. In our experiment we will only consider health profiles consisting of three periods and, hence, we will take $T=3$ in what follows. A *prospect* $(p{:}q;\ r)$ gives health profile q with probability p and health profile r with probability $1-p$.[1] Throughout the paper we will only use prospects involving at most two different health profiles.

A preference relation $\succcurlyeq$ is given over the set of prospects. The conventional notation $\succ$ and $\sim$ is used to denote strict preference and indifference. By restricting attention to *constant prospects*, i.e. prospects for which $q=r$ or for which $p=0$ or

[1] This means that there is an event E with probability p such that x obtains under E and y obtains under the complement of E. That is, we assume richness of the set of events.

1, a preference relation over health profiles can be defined, which we also denote by $\succcurlyeq$. It is implicit in the notation $(p{:}q;\ r)$ that health profile q is at least as good as health profile r ($q \succcurlyeq r$), i.e. all prospects are *rank-ordered*.

We assume that a prospect $(p{:}q;\ r)$ can be evaluated through

$$\pi U(q) + (1-\pi)U(r) \tag{1}$$

and choices and preferences correspond with this evaluation. In Eq. (1), π is the decision weight assigned to the health profile q that obtains with probability p. This decision weight is entirely general. It depends on the probability p but we assume nothing about the way in which it depends on p. We will refer to Eq. (1) as *general rank-dependent utility* (GRU). Eq. (1) is consistent with many theories of decision under risk. For example, if $\pi = p$ then Eq. (1) reduces to expected utility. If $\pi = w(p)$, with w a probability weighting function[2] then Eq. (1) reduces to rank-dependent utility (Quiggin, 1981). If $\pi = 0$ then all weight is given to the worst health profile and Eq. (1) corresponds to maximin. Eq. (1) was first suggested by Miyamoto (1988) and was subsequently used by Miyamoto and Wakker (1996) and Bleichrodt and Quiggin (1997).

Under the *QALY model* the function $U(\cdot)$ in Eq. (1) is additive:

$$U(q) = \sum_{t=1}^{T} V_t(q_t), \tag{2}$$

where the functions V_t can be period-specific. Often a more restrictive QALY model is used where the functions V_t are common for all periods and a constant discount factor is applied to all periods:

$$U(q) = \sum_{t=1}^{T} \delta^{t-1} V(q_t). \tag{3}$$

The focus in this paper is on Eq. (2), which captures the essential idea of QALYs of additivity over time. Bleichrodt and Gafni (1996) showed how Eq. (3) can be obtained from Eq. (2) by adding one preference condition.

Let $a_i v_j q$ denote the health profile q with health state q_i replaced by a_i and health state q_j replaced by v_j, $i,j \in \{1, 2, 3\}$, $i \neq j$. For example, if $i = 1, j = 2$, then $a_i v_j q = (a_1, v_2, q_3)$. Consider the following condition:

Definition 1. The preference relation $\succcurlyeq$ satisfies *generalized marginality* when for all $i,j \in \{1, 2, 3\}$, $i \neq j$, and for all health profiles q, health states a, b, c, d, v, w, x, y, and for all p:

$$(p{:}a_i v_j q;\ b_i w_j q) \sim (p{:}c_i v_j q;\ d_i w_j q) \Leftrightarrow (p{:}a_i x_j q;\ b_i y_j q) \sim (p{:}c_i x_j q;\ d_i y_j q).$$

Consider first the indifference $(p{:}a_i v_j q;\ b_i w_j q) \sim (p{:}c_i v_j q;\ d_i w_j q)$. In terms of marginal probabilities the two prospects are almost identical except that the first one gives a probability p of health state a in period i and a probability $(1-p)$ of health state b in period i and the second a probability p of health state c in period i and a probability $1-p$ of health state d in period i. Both prospects give a probability p of health state v in period j and a probability $1-p$ of health state w in period j and a probability 1 of getting q in the remaining period k. Hence, in terms of marginal probabilities the indifference implies that getting a_i with probability p and b_i with probability $1-p$ is just sufficient to offset getting c_i with probability p and d_i with probability $1-p$.

The only difference in the second indifference, $(p{:}a_i x_j q;\ b_i y_j q) \sim (p{:}c_i x_j q;\ d_i y_j q)$, is that there is a change in what happens in time period j: health state v is replaced by health state x and health state w by health state y. The latter change is such that the two prospects still yield the same marginal probability distribution over what happens in time period j: in both prospects there is a probability p of obtaining health state x in period j and a probability $1-p$ of obtaining health state y. Generalized marginality says that this change should not affect indifference. Getting a_i with probability p and b_i with probability $1-p$ should still be just sufficient to offset getting c_i with probability p and d_i with probability $1-p$. Essentially, generalized marginality says that preferences depend only on marginal probabilities (hence the term marginal in generalized marginality) and not on the joint probability distribution.

An example may clarify the restrictiveness of generalized marginality. Let there be four health states: good health, fair health, poor health, and very poor health. Suppose that a decision maker is indifferent between

$$\left(\frac{1}{2} : (\text{good, good, poor});(\text{poor, fair, poor})\right) \quad \text{and} \quad \left(\frac{1}{2} : (\text{fair, good, poor});(\text{fair, fair, poor})\right).$$

In both prospects there is a possibility of good health. In the first one the probability of good health is higher but there is also a higher probability of poor health. Generalized marginality implies that the decision maker should also be indifferent between

$$\left(\frac{1}{2} : (\text{good, poor, poor});(\text{poor, very poor, poor})\right) \quad \text{and} \quad \left(\frac{1}{2} : (\text{fair, poor, poor});(\text{fair, very poor, poor})\right).$$

[2] That is, w is increasing (if $p > q$ then $w(p) > w(q)$) and satisfies $w(0) = 0$ and $w(1) = 1$.

It is, however, conceivable that a decision maker is not indifferent between these latter two prospects. For example, he may prefer the first prospect because this gives at least some time in good health, whereas in the second prospect both health profiles are pretty bad. The example shows that there are no a priori reasons why people should or would behave according to generalized marginality.

It is easy to show that under the GRU model, the QALY model (Eq. (2)) implies generalized marginality. To improve the understanding of what generalized marginality entails, we give the proof in the main text. Let $k \neq i,j$. Under GRU and the QALY model, $(p{:}a_i v_j q; b_i w_j q) \sim (p{:}c_i w_j q; d_i x_j q)$ implies that

$$\begin{aligned}&\pi(V_i(a_i)+V_j(v_j)+V_k(q_k))+(1-\pi)(V_i(b_i)+V_j(w_j)+V_k(q_k))\\&\quad=\pi(V_i(c_i)+V_j(v_j)+V_k(q_k))+(1-\pi)(V_i(d_i)+V_j(w_j)+V_k(q_k))\end{aligned} \tag{4a}$$

or

$$\pi V_i(a_1)+(1-\pi)V_i(b_i)=\pi V_i(c_i)+(1-\pi)V_i(d_i). \tag{4b}$$

Eq. (4b) implies that

$$\begin{aligned}&\pi(V_i(a_i)+V_j(x_j)+V_k(q_k))+(1-\pi)(V_i(b_i)+V_j(y_j)+V_k(q_k))\\&\quad=\pi(V_i(c_i)+V_j(x_j)+V_k(q_k))+(1-\pi)(V_i(d_i)+V_j(y_j)+V_k(q_k)).\end{aligned}$$

and substitution of $\pi V_i(a_1)+(1-\pi)V_i(b_i)=\pi V_i(c_i)+(1-\pi)V_i(d_i)$ implies $(p{:}a_i x_j q;\ b_i y_j q) \sim (p{:}c_i x_j q;\ d_i y_j q)$.

Bleichrodt and Quiggin (1997, Theorem 4) showed that under GRU, the QALY model not only implies generalized marginality, but generalized marginality also implies the QALY model.[3] Hence, generalized marginality is the central condition of the QALY model.

We next define utility independence. Let J be a subset of $\{1, 2, 3\}$ and let q and k be two health profiles. By $k_J q$ we denote the health profile q with health state q_j replaced by health state k_j for all j in J. For example, if $J=\{1, 3\}$ then $k_J q=(k_1, q_2, k_3)$.

Definition 2. The preference relation $\succcurlyeq$ satisfies *utility independence* if for all subsets J of $\{1, 2, 3\}$, for all health profiles k, l, m, n, q, r, and for all probabilities p:

$$(p : k_J q; l_J q) \sim (p : m_J q; n_J q) \Leftrightarrow (p : k_J r; l_J r) \sim (p : m_J r; n_J r).$$

That is, if all health profiles in the prospects under comparison have common health states outside J preferences do not depend on what these common health states are.

Consider, again, the example given before. If the decision maker is indifferent between

$$\left(\frac{1}{2} : (\text{good, good, poor}); (\text{poor, fair, poor})\right) \quad \text{and} \quad \left(\frac{1}{2} : (\text{fair, good, poor}); (\text{fair, fair, poor})\right)$$

then utility independence says that he should also be indifferent between

$$\left(\frac{1}{2} : (\text{good, good, good}); (\text{poor, fair, good})\right) \quad \text{and} \quad \left(\frac{1}{2} : (\text{fair, good, good}); (\text{fair, fair, good})\right),$$

where we changed the common outcome in the third period from poor to good. It is, however, conceivable that the decision maker is not indifferent between the second pair of prospects. He may now, for instance, prefer the second prospect because it does not carry the risk of spending time in poor health. Like generalized marginality, utility independence is not a priori obviously fulfilled.

Utility independence is less restrictive than generalized marginality: generalized marginality implies utility independence but the reverse is not true. Miyamoto and Wakker (1996, Theorem 4) showed that if utility independence holds but generalized marginality is violated then

$$U(q)=\prod_{t=1}^{T} V_t(q_t), \tag{5}$$

i.e. utility is multiplicative. Eq. (5) is still tractable. Consequently, not all is lost when generalized marginality is violated and, in the face of possible violations of generalized marginality, it is important to test utility independence.

Guerrero and Herrero (2005) further relaxed utility independence and only imposed it for initial health states. They showed that even then a reasonably tractable model results. Their condition is hard to test empirically because it involves dynamic decisions and tests of their model require the comparison of choices made at different points in time. We do not consider their model in this paper.

[3] This result requires richness of the set of health states, i.e. indifferences can be obtained by variations in the health states. For empirical testing it is easier to vary probabilities and, therefore, we assumed a rich set of events (see footnote 1). Whether the presence of a rich set of events implies that the result of Bleichrodt and Quiggin (1997) still holds without richness of the set of health states assumed is an open question.

3. Experiment

3.1. Test

The aim of the experiment was to test generalized marginality and utility independence. The general structure of our tests of generalized marginality was as follows. First we elicited the probability p_{vw} such that a subject was indifferent between prospects of the type $(p_{vw} : a_i v_j q; b_i w_j q)$ and $(p_{vw} : c_i v_j q; d_i w_j q)$. Then we elicited the probability p_{xy} such that subjects were indifferent between $(p_{xy} : a_i x_j q; b_i y_j q)$ and $(p_{xy} : c_i x_j q; d_i y_j q)$ with x and y different from v and w. Under generalized marginality we should observe that $p_{vw} = p_{xy}$ except for random error.

To test for utility independence we first elicited the probability p_q such that subjects were indifferent between $(p_q{:}k_J q;\ l_J q)$ and $(p_q{:}m_J q;\ n_J q)$ for a given subset J. Then we elicited the probability p_r such that subjects were indifferent between $(p_r{:}k_J r;\ l_J r)$ and $(p_r{:}m_J r;\ n_J r)$ with r different from q. Under utility independence we should observe that $p_q = p_r$ except for random error.

3.2. Subjects

Subjects were 60 students (30 female, median age of all subjects 22 years) from Erasmus University. They were paid a flat fee of €10. Prior to the actual experiment, the experimental design was tested and fine-tuned in several pilot sessions.

3.3. Procedure

The experiment was run on a computer in personal interview sessions. To reduce errors and to ensure that subjects could focus entirely on the experimental questions, all responses were entered into the computer by the interviewer. Subjects were told that there were no right or wrong answers and that we were only interested in their preferences. Experimental sessions lasted 40 min on average and consisted of three parts: instructions and practice questions, data collection for the experiment reported here, and data collection for an unrelated experiment. Subjects took approximately 15–20 min to answer the questions for this experiment.

All indifference probabilities were elicited through a series of choices. Each choice question corresponded to an iteration in a bisection process, which is described in Appendix B. A choice-based elicitation procedure was used because previous studies observed that inferring indifferences from a series of choices leads to fewer inconsistencies than asking subjects directly for their indifference value (see Luce (2000) for a review). The iterative process ended when the absolute difference in probability between successive steps in the iteration was less than 5%. We learned from the pilot sessions that it was unrealistic to determine the probabilities with more precision. At the end of each iteration process we repeated the first question of the iteration process. If subjects gave the same answer to this repeated choice question then we moved on to the next elicitation. If not, the iteration process for this elicitation was started anew. The aim of repeating the first choice in the iteration process was to reduce the impact of decision errors.

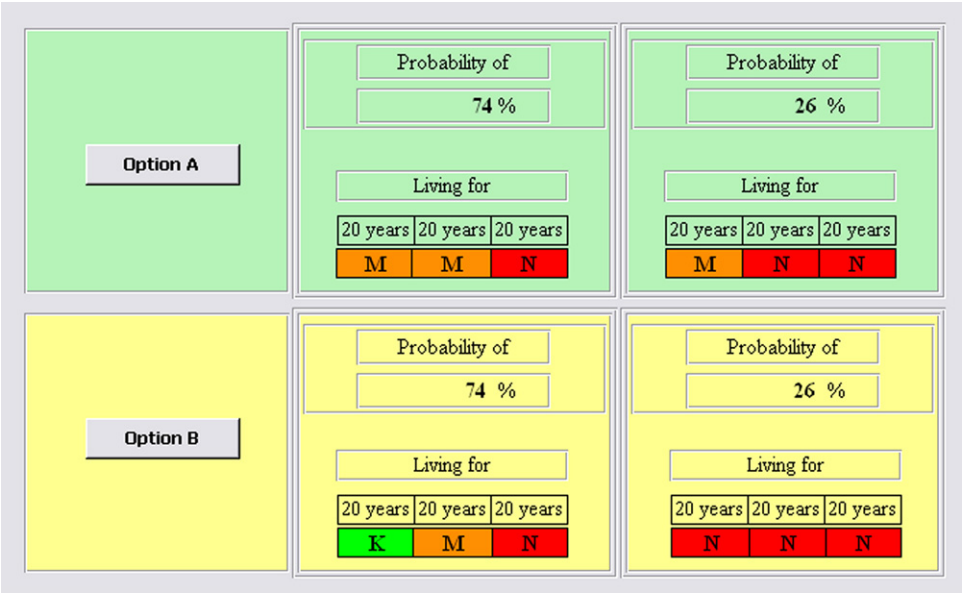


Fig. 1. Example of an experimental question.

Table 1
Description of health states used in the experiment

Label	Color	EQ code	EQ utility
K	Green	11111	1.000
L	Yellow	11121	0.850
M	Orange	11122	0.722
N	Red	12222	0.551

Table 2
Tests of generalized marginality

Test	Part	Question
1	I	(p:MMN; MNN) vs. (p:KMN; NNN)
	II	(p:MKN; MMN) vs. (p:KKN; NMN)
2	I	(p:KLM; KMN) vs. (p:KKM; KNN)
	II	(p:KLL; KMM) vs. (p:KKL; KNM)
3	I	(p:KLM; LLM) vs. (p:KLL; LLN)
	II	(p:KLM; MLL) vs. (p:KLL; MLN)

3.4. Stimuli

Subjects were asked to make a choice between two prospects consisting of two health profiles, neutrally labeled A and B. Fig. 1 shows the way the prospects were displayed on the computer screen.

Health profiles consisted of three periods of 20 years each. Hence, the total length of the profiles was 60 years, which corresponded to the life-expectancy of our subjects. We used only three periods to keep the tasks as simple as possible. We used periods of 20 years because these more or less correspond to different life stages. The health states constituting the health profiles were selected from a set of four EuroQol health states. We selected only moderate health states because these can be imagined more easily by a healthy population like our subjects. Another reason to use moderate health states was to avoid considerations of maximal endurable time (Stalmeier et al., 1997). Health states were labeled using capital letters from the middle of the alphabet, minimizing potential distorting associations (for example using the letter D might be associated with the outcome death). The ordering of the health states was obvious in the sense that more preferred health states scored at least as good on each EuroQol dimension as less preferred health states and strictly better on at least one dimension. The ordering of the health states corresponded with the alphabetical order.

Health states were printed on separate cards and were assigned colors in an intuitive order (green the best, red the worst, etc.). The use of color-coding aimed to facilitate decision-making by reminding the subjects of the relative attractiveness of the health states. The EuroQoL system was introduced in the initial instructions and, throughout the experiment, subjects had the cards describing the health states in front of them. Health states are summarized in Table 1 and the cards that were handed to the subjects are reproduced in Appendix A. The final column of Table 1 displays the utility of the health states according to the EuroQol algorithm (Dolan, 1997).

It is crucial for our tests that the prospects are rank-ordered. To ensure this we selected the prospects such that one profile yielded in each period a health state that was always at least as good as the other profile. To help subjects understand the ranking of health profiles in each of the choices they faced, we asked them – before the bisection procedure for a particular question started – to rank the four health profiles involved in the question from the best to the worst with ties allowed. No violations of rank-ordering were observed.

We performed three tests of generalized marginality and four tests of utility independence. The tests of generalized marginality are displayed in Table 2, those of utility independence in Table 3. All tests consisted of two parts. The first part of a test is indicated with the Roman number I in the tables, the second part with the number II. The prospect mentioned first was displayed to the subjects as option A, the other as option B. KMN denotes a profile that gives health state K for the first

Table 3
Tests of utility independence

Test	Part	Question
1	I	(p:LLM; MNM) vs. (p:LMM; MMM)
	II	(p:LLN; MNN) vs. (p:LMN; MMN)
2	I	(p:KKL; KNN) vs. (p:KML; KMN)
	II	(p:LKL; LNN) vs. (p:LML; LMN)
3	I	(p:LKN; LNN) vs. (p:LMN; LMN)
	II	(p:NKN; NNN) vs. (p:NMN; NMN)
4	I	(p:KML; KMN) vs. (p:KMM; KMM)
	II	(p:MLL; MLN) vs. (p:MLM; MLM)

20 years, health state M for the next 20 years and health state N for the final 20 years. Outcomes that were varied between the two parts of each test are underlined. Note that in the fourth test of utility independence two common outcomes were changed.

The order in which the questions were asked was arbitrary with the restriction that the two parts of a given test were never offered consecutively. Interspersing trials were implemented to prevent subjects from forming a match that would guide answers.

The experiment ended with two consistency tests in which subjects repeated the first part of the first test of generalized marginality (GM1-I) and the second part of the third test of utility independence (UI3-II).

Spencer and Robinson (2007) also tested utility independence and found support for it in six out of eight tests. The main difference between their study and ours is that they asked directly for indifferences whereas we used a choice-based elicitation method.[4] It is well known from the literature that matching and choice invoke different cognitive processes (Tversky et al., 1988). If the two studies were to give similar results in spite of the different response modes used then this would offer convincing evidence in favor of utility independence.

3.5. Analysis

We used both parametric (t-test) and nonparametric (Wilcoxon) tests to test for significance of differences. Unless a difference was observed we only report the parametric results. Because we performed many different tests, there is a danger of finding significant differences just by chance. To reduce this danger we used a significance level of 1% in the statistical tests reported below. Using 5% instead did not affect our conclusions much.

Our sample size with a standard deviation of 0.15 would have enough power to detect a difference in aggregate values of 0.08 ($\alpha = 0.01$, $1 - \beta = 0.90$). In the literature a difference of 0.10 is often considered meaningful and important in decision-making contexts (O'Brien and Drummond, 1994). Hence, the power of our study was satisfactory.

4. Results

4.1. Consistency

Three subjects were excluded either for not cooperating or for targeting towards 0% and 100% in each question. This left 57 subjects in the final analysis. The consistency tests yielded mixed results. In the test for GM1-I the median probabilities were 0.54 in the original test and 0.58 in the retest. The difference was not significant ($p = 0.064$). The median of the individual absolute differences between test and retest was 0.06. In the test for UI3-II the median probabilities were 0.62 in the original experiment and 0.60 in the retest. In this case, the difference was significant, however (t-test, $p = 0.005$ and Wilcoxon test, $p = 0.011$). The median of the individual absolute differences between test and retest was 0.04. The Pearson correlation coefficients between test and retest were in both cases 0.63.

Few subjects had to restart the iteration process because they reversed their first choice. For the median subject this happened in just 1 out of 16 tests. In total, the proportion of reversals was 11.3%. This suggests that errors were rare when probabilities differed substantially from their indifference values, which was usually the case in the initial choices. By comparison, reversals up to 33% are common in choice experiments (Stott, 2006).

4.2. Aggregate results

Fig. 2 displays the medians of the elicited indifference probabilities. The means were similar. The figure shows support for generalized marginality: for all three tests the median probabilities for both parts of the test were very close. None of the differences was significant ($p > 0.60$ in all three tests). The correlation coefficients between the two parts were, however, rather low. Correlation was only fair for GM1 (0.27) and GM2 (0.30) and was moderate (0.52) for GM3.

Fig. 2 shows that the differences between the median probabilities were generally larger in the tests of utility independence than in the tests of generalized marginality. There appears to be no systematic pattern in the medians, however. In the first test of utility independence (UI1), the indifference probability was larger in the second part, in UI2 and UI3 it was larger in the first part and in UI4 there was no difference. We could not reject utility independence in three out of four tests (UI1, UI3, and UI4). The only exception is the second test. Here the difference in elicited probabilities was significant (t-test, $p = 0.008$ and Wilcoxon test, $p = 0.016$). Correlations between the two parts of the tests are higher than for generalized marginality and are moderate in all tests (0.44 in UI1, 0.46 in UI2, 0.51 in UI3, and 0.55 in UI4).

[4] Other differences are that they used group sessions of 10–20 subjects whereas we used personal interviews, they used pen and paper whereas our experiment was computer-run, they used 5 periods of 5 years whereas we used three periods of 20 years, and they used health states ranging from normal health to death whereas we used only moderate health states. Finally, they asked the two parts of the tests of utility independence consecutively in their first experiment, but not in their second experiment, which randomized the three tests that were used.

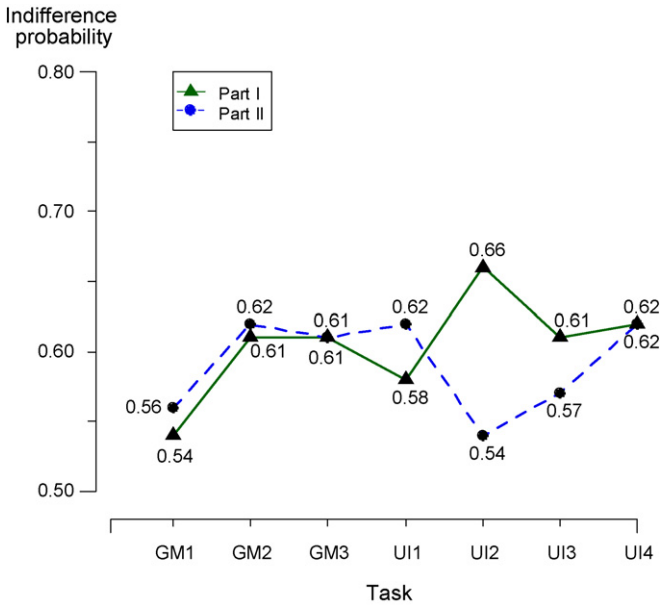


Fig. 2. Median probabilities in the two parts of the tests.

4.3. Individual results

Fig. 3 shows the means and the medians of the individual absolute differences between the elicited probabilities in the two parts of each test. It should be kept in mind when interpreting these results that our choice-based procedure was terminated when the absolute difference in probabilities between successive iterations was less than 0.05 and the indifference value was set equal to the midpoint of the elicited interval (see Appendix B for details). This implies that there could be a maximum difference of 0.04 between the elicited probability and the true probability. Hence, when we compare the probabilities between the two parts of a test a difference of 0.08 might in theory have been caused by our elicitation procedure.

Table 4 shows the probabilities of observing a given absolute difference under our elicitation method when in reality no difference exists. The table was constructed under the assumption that any value from the elicited indifference interval was as likely to be the true indifference value. Table 4 displays that observing a difference of 0.08 due to our elicitation method alone was very unlikely: the chance of it happening was less than 0.001. The chance that our elicita-

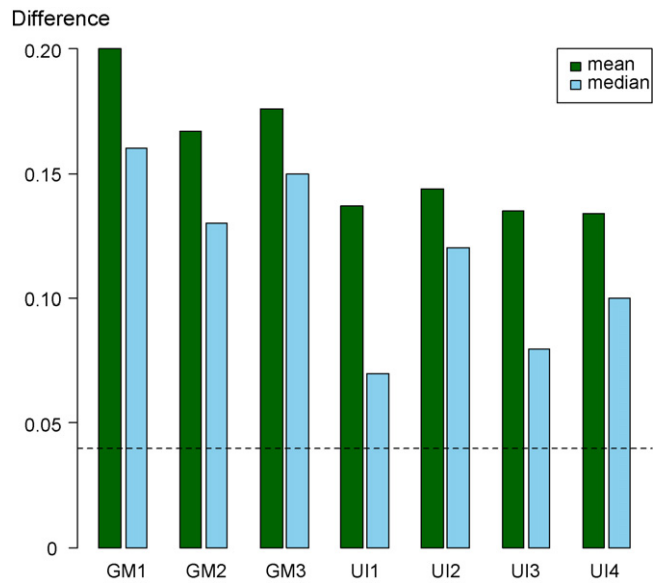


Fig. 3. Mean and median absolute difference between the two parts of the test.

Table 4
Probability of a difference in probability being caused by elicitation method

Difference	Probability
0	0.148
0.01	0.276
0.02	0.226
0.03	0.166
0.04	0.106
0.05	0.053
0.06	0.019
0.07	0.004
0.08	<0.001

Table 5
Classification of subjects based on the number of times they violated generalized marginality using different thresholds

Number of violations	Threshold	
	0.08	0.13
0	4	12
1	12	13
2	22	21
3	19	11

tion method led to an observed difference larger than 0.04 was, in fact, only 0.077. Differences up to 0.04 were, however, plausible. To illustrate this, we have plotted the value of 0.04 by a dotted line in Fig. 3. Up to this line differences might reasonably be attributed to our elicitation method. Above it they cannot be explained by our elicitation method alone. The figure shows that both for utility independence and, in particular, for generalized marginality the observed absolute differences are clearly larger than 0.04 and, hence, they are not just products of imprecision in our elicitation method.

Another thing to take into account when considering Fig. 3 is that subjects' preferences are likely to be imprecise (Dubourg et al., 1994; Butler and Loomes, 2007). The choices we asked our subjects to make were not easy and subjects had to compare probabilities, health states, and the timing and sequence of the health states simultaneously. When faced with complex choices it seems unrealistic to expect that subjects always have clear preferences between the two options. If subjects were, for instance, only able to distinguish between probabilities when they differed by at least 0.05 then an observed difference of 0.14 between the two parts of the tests could in theory be entirely caused by our elicitation method and preference imprecision. This value was however extremely unlikely: it had a probability of 0.0002. Of course we do not know exactly how much of the differences that we observed were actually caused by preference imprecision but it is likely to have played at least some role in the observed differences. The medians of the individual differences between the test and retest for GM1-I and UI3-II may give some indication of this imprecision. They were 0.06 and 0.04 respectively and the value of 0.05 that we used in the above example was selected because it is the midpoint of these two values. It was also the median imprecision that was observed by Bleichrodt and Johannesson (1997) who made an attempt to quantify preference imprecision in health utility measurement.

Table 5 presents a classification of our subjects based on the number of times that their responses exceeded a given threshold in the tests of generalized marginality. To account for the possibility of differences due to the elicitation procedure and due to preference imprecision we report the results for two different thresholds: 0.08 and 0.13. The table shows, for example, that there were only four subjects for whom the difference between the two parts of the tests was less than 0.08 in all three tests of generalized marginality. The conclusions drawn from the table depend on the threshold that is deemed plausible. Regardless of the threshold used, it seems safe to conclude that a substantial proportion of our subjects violated generalized marginality and, consequently, the QALY model.

Table 6 presents the same classification for the tests of utility independence. A comparison between Tables 5 and 6 reveals that there is more support for utility independence than for generalized marginality at the individual level. If we use a threshold of 0.13, over 60% of the subjects satisfied utility independence in at least three out of four tests. There were

Table 6
Classification of subjects based on the number of times they violated utility independence using different thresholds

Number of violations	Threshold	
	0.08	0.13
0	4	14
1	16	21
2	23	15
3	12	7
4	2	0

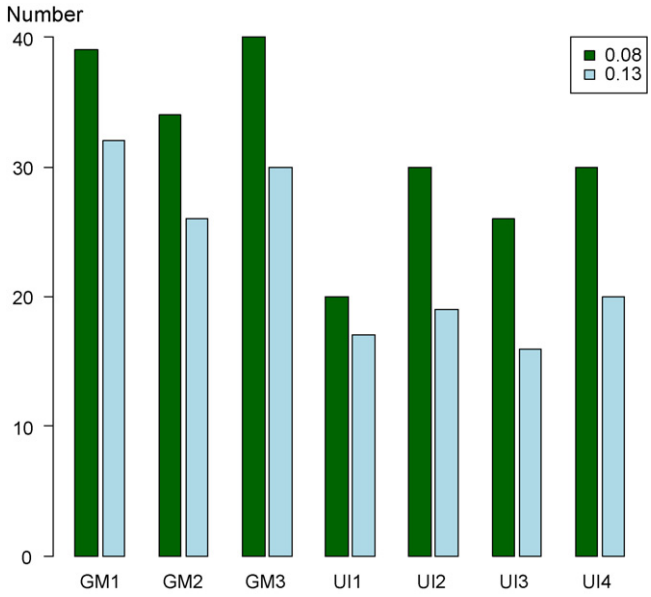


Fig. 4. Number of subjects violating a test for two different thresholds.

hardly any subjects who violated utility independence in each test, whereas the proportion of subjects violating generalized marginality in each test was substantial.

Fig. 4 shows the number of subjects violating a particular test. The figure confirms that violations were more common in the tests of generalized marginality than in the tests of utility independence. The figure also shows that violations were not confined to one particular test but occurred in all tests.

5. Discussion

5.1. Main findings

We have performed new tests of the QALY model and a generalization thereof. Our tests do not require additional confounding assumptions, for example about discounting, and take account of violations of expected utility. At the aggregate level we observed support for the QALY model as we could not reject generalized marginality, the central condition of the QALY model. At the individual level there is much less support for the QALY model: a sizeable proportion of our subjects violated generalized marginality and the observed deviations were too large to be caused by elicitation and preference imprecisions alone.

We also tested for utility independence, a less restrictive preference condition than generalized marginality, which still implies a tractable model. Utility independence was supported at the aggregate level. At the individual level we found more support for utility independence than for generalized marginality. For a substantial proportion of our subjects the observed deviations from utility independence can reasonably be attributed to the elicitation procedure and preference imprecision. Our aggregate findings on utility independence are consistent with the findings of Spencer and Robinson (2007) in spite of the differences in response mode and experimental design between their and our study. The data in Spencer and Robinson (2007) and in our study reinforce each other and provide a strong case for utility independence at the aggregate level. Spencer and Robinson do not report individual-level results.

5.2. Caveats

The decision tasks used in our experiment were cognitively demanding. Subjects had to take into consideration several dimensions simultaneously (probability, quality of life, and duration and sequence of the health states). We took several precautions to try and keep the experimental tasks as simple as possible by using just four easily imaginable color-coded health states, by using only three time periods of equal length, and by using a computer-run choice-based questionnaire. Nevertheless, subjects may have adopted simplifying heuristics to facilitate responding. Two such heuristics might a priori be particularly plausible.

First, subjects may have made the tasks easier by targeting towards probability 0.50. We had no indication that subjects indeed used this heuristic. First, most elicited probabilities differed significantly from 0.50. Second, there was no subject for whom all elicited indifference probabilities fell between 0.40 and 0.60. The data do not suggest that subjects were

targeting towards another probability either. When we compared differences in elicited probabilities across unrelated decision tasks (e.g. compare GM1-I with GM3-II) then many significant differences were observed. The latter observation also shows that the fact that we could not reject generalized marginality and utility independence at the aggregate level was not due to a lack of power in our study. Empirically meaningful differences in elicited probabilities can be elicited in our sample.

A second heuristic that subjects could have employed in the utility independence questions was to cancel out the common health states. For example, in the comparison between (p:LLM; MNM) and (p:LMM; MMM), task UI1-I, subjects may have made the task easier by eliminating the common health state M in the third period. Adopting such a strategy would make the two parts of each test of utility independence identical and would create artificial support for utility independence. The experimental design took care to avoid that subjects would use this heuristic. In particular, we randomized the order of the tests so that subjects never faced the two parts of a test consecutively. It cannot be excluded though that at least some subjects adopted this cancellation heuristic in spite of the precautions we took.

We used students as subjects. We do not believe that this limits the generalizability of our findings. Empirical evidence on health utility measurement has shown that there exist no significant differences between the patterns of responses obtained from convenience samples and those obtained from representative samples from the general population. For a review see de Wit et al. (2000) and for a more recent comparison Bleichrodt et al. (2005).

Our results show that many subjects deviate from generalized marginality casting doubt on the descriptive appeal of the QALY model. These results say nothing about the normative validity of generalized marginality. One might argue that it is desirable for normative reasons to accept generalized marginality and interpret the deviations that we observed as irrationalities that reflect biases in time aggregation that we should seek to correct. We do not agree with this view. We do not consider generalized marginality normative and, as we explained in Section 2, there are good reasons why people may deviate from it.

We implicitly assumed that QALYs should reflect individual preferences for health. There is an alternative, extra-welfarist, strand in the literature, which takes QALYs as a measure of health and not necessarily as a reflection of people's preferences for health. The two approaches are not necessarily incompatible. Extra-welfarists use preference-based quality weights to quantify QALYs which suggests that even in the extra-welfarist approach individual preferences are important. Further, even if one takes the position that QALY are only a measure of health then one would expect that people's preferences are increasing in QALYs (better health is desirable). Given the uniqueness properties of utility, this essentially means that QALYs should reflect individual preferences. Hence, even in the extra-welfarist approach people's preferences and consequently our tests are important.

Our tests are robust to many violations of expected utility. As we explained in Section 2, the utility model that we assumed includes as special cases many of the theories of decision under risk that are available today. Hence, our results are not affected by the deviations from expected utility modeled by these theories. This is not to say that our results are robust to all deviations from expected utility. Our results depend on the validity of the general utility model and, even though very general, the model is not consistent with any preference pattern. For instance, the model makes no distinction between gains and losses and, hence, it is not robust to loss aversion. Developing tests that are robust to loss aversion is an important challenge. That said, our method corrects for more biases than any previous study and, hence, our tests are the most powerful tests of the QALY model available today.

5.3. Implications

Our results provide support for the QALY model at the aggregate level. It should be pointed out though that this conclusion is based on three tests only. It should also be kept in mind that we only used mild to moderate health states to avoid considerations like maximal endurable time. Our conclusions may no longer hold when more severe health states are involved. More evidence is needed and we invite other researchers to try and replicate our findings using other experimental designs.

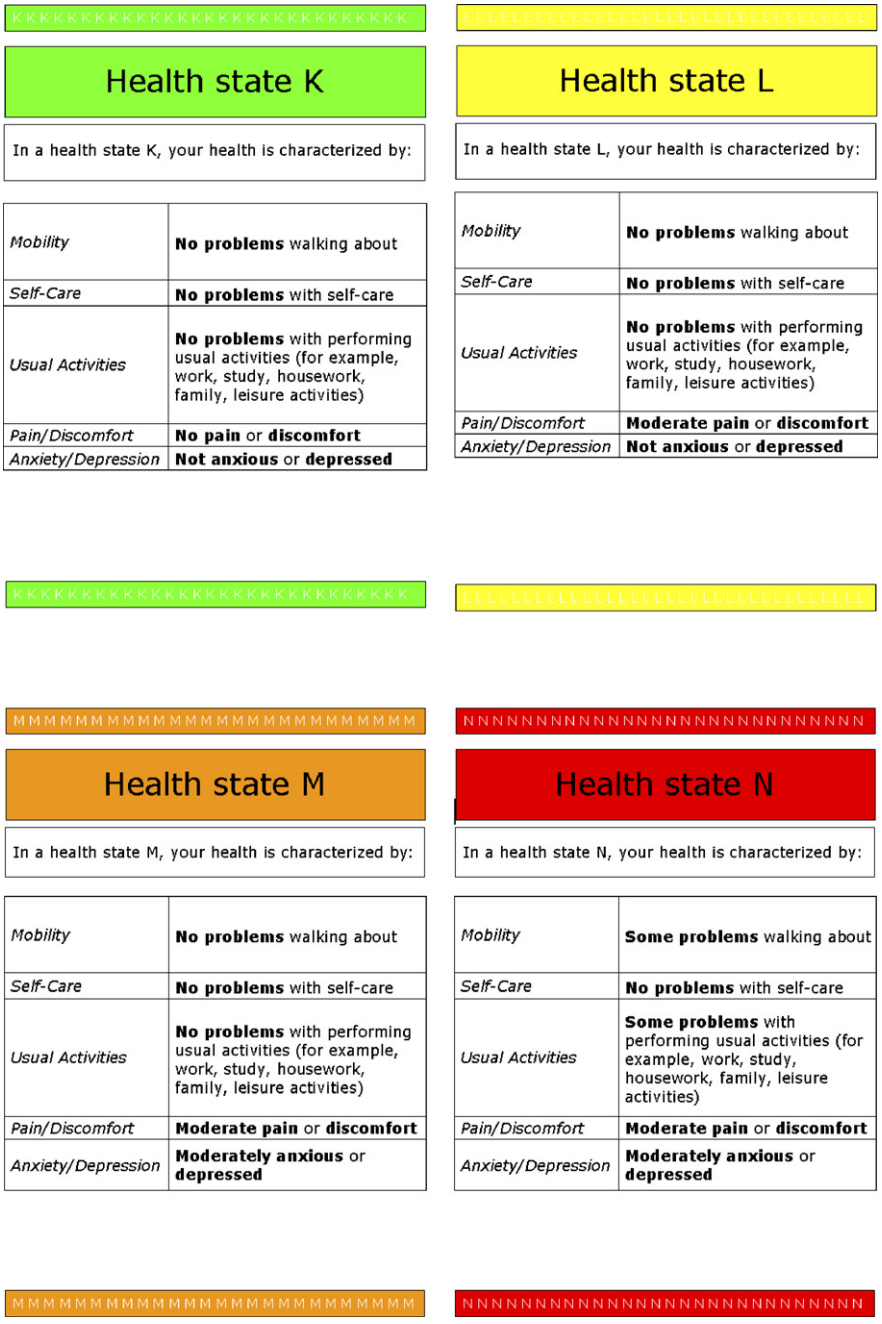
At the individual level, the support for QALYs appears weaker. Our data suggest that QALYs cannot be applied in individual medical decision-making without some additional tests of the decision maker's preference structure. The tests developed in this paper may be helpful in doing so. In interpreting our results at the individual level, one should keep in mind though that the tasks were demanding and that there was a possibility of substantial imprecision in subjects' responses. Hence, this single study should not be taken as conclusive evidence of the validity of QALYs at the individual level.

Even when QALYs are found not to hold, not all is lost. Our results, suggest that there is more support for utility independence at the individual level. Utility independence still implies a tractable model that can be applied in practice. Hence, in contrast with a frequently voiced belief that QALYs are not consistent with people's preferences for health, the overall message of this paper is supportive of the use of QALY-type models in health economics.

Acknowledgements

We are grateful to Anirban Basu, Willard Manning, Elly Stolk, Peter Wakker, and three anonymous referees for helpful comments on a previous version of this paper. Han Bleichrodt's research was supported by a research grant from the Netherlands Organization for Scientific Research (NWO). Martin Filko's research was supported by a grant from DSW.

Appendix A. Description of health states used in the experiment



Health state K

In a health state K, your health is characterized by:

Mobility	**No problems** walking about
Self-Care	**No problems** with self-care
Usual Activities	**No problems** with performing usual activities (for example, work, study, housework, family, leisure activities)
Pain/Discomfort	**No pain** or **discomfort**
Anxiety/Depression	**Not anxious** or **depressed**

Health state L

In a health state L, your health is characterized by:

Mobility	**No problems** walking about
Self-Care	**No problems** with self-care
Usual Activities	**No problems** with performing usual activities (for example, work, study, housework, family, leisure activities)
Pain/Discomfort	**Moderate pain** or **discomfort**
Anxiety/Depression	**Not anxious** or **depressed**

Health state M

In a health state M, your health is characterized by:

Mobility	**No problems** walking about
Self-Care	**No problems** with self-care
Usual Activities	**No problems** with performing usual activities (for example, work, study, housework, family, leisure activities)
Pain/Discomfort	**Moderate pain** or **discomfort**
Anxiety/Depression	**Moderately anxious** or **depressed**

Health state N

In a health state N, your health is characterized by:

Mobility	**Some problems** walking about
Self-Care	**No problems** with self-care
Usual Activities	**Some problems** with performing usual activities (for example, work, study, housework, family, leisure activities)
Pain/Discomfort	**Moderate pain** or **discomfort**
Anxiety/Depression	**Moderately anxious** or **depressed**

Appendix B. Explanation of the bisection method

The bisection method used to generate the iterations is illustrated in Table B1 for task GM1I. The option that is chosen is printed in bold. The starting probability in the iterations was determined randomly. Depending on the choice made, the probability was increased or decreased. The size of change in the second iteration was half the difference between the probability in the first question and 0 or half the difference between the probability in the first question and 1. Which one was selected depended on the subject's choice. The size of the change in the remaining iterations was half the size of the change in the previous question. The iteration process ended when the difference between the probability and the previous probability was less than 0.05. The iteration process resulted in an interval within which the indifference value should lie. The midpoint of this interval was taken as the indifference value. For example, in Table B1 the indifference value for p should lie between 0.63 and 0.68. Then we took as the indifference value 0.66.

Table B1
An illustration of the bisection method

Iteration	Offered choices
1	(0.72:MMN; MNN) ∼ (**0.72:KMN; NNN**)
2	(**0.36:MMN; MNN**) ∼ (0.36:KMN; NNN)
3	(**0.54:MMN; MNN**) ∼ (0.54:KMN; NNN)
4	(**0.63:MMN; MNN**) ∼ (0.63:KMN; NNN)
5	(0.68:MMN; MNN) ∼ (**0.68:KMN; NNN**)
Indifference value	0.66

References

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., Doctor, J.N., Stolk, E.A., 2005. A nonparametric elicitation of the equity-efficiency trade-off in cost-utility analysis. Journal of Health Economics 24, 655–678.

Bleichrodt, H., Gafni, A., 1996. Time preference, the discounted utility model and health. Journal of Health Economics 15, 49–66.

Bleichrodt, H., Johannesson, M., 1997. The validity of QALYs: an empirical test of constant proportional tradeoff and utility independence. Medical Decision Making 17, 21–32.

Bleichrodt, H., Pinto-Prades, J.-L., 2006. Conceptual foundations for health utility measurement. In: Jones, A.M. (Ed.), The Elgar Companion to Health Economics. Edward Elgar, Aldershot, pp. 347–358.

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

Brazier, J., Dolan, P., Karampela, K., Towers, I., 2006. Does the whole equal the sum of the parts? Patient-assigned utility scores for ibs-related health states and profiles. Health Economics 15, 543–551.

Butler, D., Loomes, G., 2007. Imprecision as an account of the preference reversal phenomenon. American Economic Review 97, 277–297.

de Wit, G.A., van Busschbach, J.J., de Charro, F.T., 2000. Sensitivity and perspective in the valuation of health status. Health Economics 9, 109–126.

Doctor, J.N., Bleichrodt, H., Lin, J.H., in press. Health utility bias: a meta-analytic evaluation. Medical Decision Making.

Doctor, J.N., Bleichrodt, H., Miyamoto, J., Temkin, N.R., Dikmen, S., 2004. A new and more robust test of QALYs. Journal of Health Economics 23, 353–367.

Dolan, P., 1997. Modeling valuations for Euroqol health states. Medical Care 35, 1095–1108.

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

Frederick, S., Loewenstein, G.F., O'Donoghue, T., 2002. Time discounting and time preference: a critical review. Journal of Economic Literature 40, 351–401.

Guerrero, A.M., Herrero, C., 2005. A semi-separable utility function for health profiles. Journal of Health Economics 24, 33–54.

Krabbe, P.F.M., Bonsel, G.J., 1998. Sequence effects, health profiles, and the QALY model: in search of realistic modeling. Medical Decision Making 18, 178–186.

Kuppermann, M., Shiboski, S., Feeny, D., Elkin, E.P., Washington, A.E., 1997. Can preference scores for discrete states be used to derive preference scores for entire paths of events? Medical Decision Making 17, 42–55.

Luce, R.D., 2000. Utility of Gains and Losses: Measurement-Theoretical and Experimental approaches. Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey.

Mackeigan, L.D., O'Brien, B.J., Oh, P.I., 1999. Holistic versus composite preferences for lifetime treatment sequences for type 2 diabetes. Medical Decision Making 19, 113–121.

Miyamoto, J.M., 1988. Generic utility theory: measurement foundations and applications in multiattribute utility theory. Journal of Mathematical Psychology 32, 357–404.

Miyamoto, J.M., Wakker, P.P., 1996. Multiattribute utility theory without expected utility foundations. Operations Research 44, 313–326.

O'Brien, B.J., Drummond, M.F., 1994. Statistical versus quantitative significance in the socioeconomic evaluation of medicines. PharmacoEconomics 5, 389–398.

Quiggin, J., 1981. Risk perception and risk aversion among Australian farmers. Australian Journal of Agricultural Economics 25, 160–169.

Richardson, J., Hall, J., Salkeld, G., 1996. The measurement of utility in multiphase health states. International Journal of Technology Assessment in Health Care 12, 151–162.

Spencer, A., 2003. A test of the QALY model when health varies over time. Social Science and Medicine 57, 1697–1706.

Spencer, A., Robinson, A., 2007. Test of utility independence when health varies over time. Journal of Health Economics 26, 1003–1013.

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.

Stott, H.P., 2006. Cumulative prospect theory's functional menagerie. Journal of Risk and Uncertainty 32, 101–130.

Treadwell, J.R., 1998. Tests of preferential independence in the QALY model. Medical Decision Making 18, 418–428.

Tversky, A., Sattath, S., Slovic, P., 1988. Contingent weighting in judgment and choice. Psychological Review 95, 371–384.

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., 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.