The Validity of QALYs:

An Experimental Test of Constant Proportional Tradeoff and Utility Independence

HAN BLEICHRODT, PhD, MAGNUS JOHANNESSON, PhD

Pliskin, Shepard, and Weinstein identified three preference conditions that ensure that quality-adjusted life years (QALYs) represent preferences over gambles over chronic health profiles. This paper presents an experimental test of the descriptive validity of two of these preference assumptions: utility independence and constant proportional tradeoff. Eighty students at the Stockholm School of Economics and 92 students at Erasmus University Rotterdam participated in the experiment. The results of the experiment support the descriptive validity of constant proportional tradeoff: both within groups and between groups constant proportional tradeoff could not be rejected. The results are less supportive of the.descriptive validity of utility independence. Within-groups utility independence was rejected. Between-groups utility independence could not be rejected, but this may have been due to a lack of statistical power. Analysis of the individual responses revealed that without adjustment for imprecision of preference, 39 respondents (22.8%) satisfied constant proportional tradeoff. Twenty-three respondents (13.4%) satisfied utility independence without adjustment for imprecision of preference. However, because of the relative unfamiliarity of the respondents with both the health states to be evaluated and the methods of health-state-utility measurement, it is likely that the respondents' preferences were imprecise. Adjusted for imprecision of preference, the upper estimates of the proportions of respondents who satisfied constant proportional tradeoff and utility independence, respectively, were 90.1% (155 respondents) and 75.8% (130 respondents). Pliskin et al. further derived that if an individual's preferences satisfy both constant proportional tradeoff and utility independence, then these preferences can be represented by a more general, risk-adjusted QALY model. Without adjustment for imprecision of preference, ten respondents (5.8%) satisfied both constant proportional tradeoff and utility independence. Adjusted for imprecision of preference, the upper estimate of the proportion of respondents who satisfied both constant proportional tradeoff and utility independence was 88.8% (118 respondents). The results of this study indicate that constant proportional tradeoff holds approximately. The evidence is much weaker for utility independence, however. This has important implications for the use of QALY-type measures in medical decision making. Key words: QALYs; health utility measurements; medical decision making: individual preferences. (Med Decis Making 1996;17:21-32)

In health care, as in other areas of social policy, decisions have to be made with respect to the allocation of scarce resources. Cost-utiliy analysis, in which quality-adjusted life years (QALYs) are used as the outcome measure, is intended to guide health-care policy making. Over the past decade, QALY-based decision making has become increasingly popular. QALYs provide a straightforward way to combine the two main outcomes of health-care programs, quantity of life and quality of life, into a single index measure. A further advantage of using QALYs is that they have intuitive appeal. However, ever since the introduction of QALYs, their theoretical properties have been a matter of concern. Pliskin, Shepard, and Weinstein' were the first to provide an axiomatic analysis of QALYs. These authors show that, given an individual preference relation over gambles involving quantity of life and constant quality of life that satisfies the axioms of expected utility theory, three conditions have to be imposed on this preference relation to ensure that it can be represented by the QALY model. These conditions are referred to as "(mutual) utility independence," "constant proportional tradeoff," and "risk neutrality on life years." Pliskin et al. have further derived that

Received April 17, 1995, from the Institute for Medical Technology Assessment, Department of Health Policy and Management, Erasmus University, Rotterdam, The Netherlands (HB); and the Centre for Health Economics, Stockholm School of Economics, Stockholm, Sweden (MJ). Revision accepted for publication February 27,1996. Supported in part by the National Corporation of Swedish Pharmacies and by Merck, Sharp and Dohme.

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

imposing utility independence and constant proportional tradeoff, but not risk neutrality on life years, ensures that the individual preference relation can be represented by a general QALY model in which life years do not enter linearly, but are adjusted for risk attitude.

Identifying the preference conditions on which the QALY model depends allows an assessment of both the extent to which it is rational for an individual to behave according to the model (i.e., the normative validity of the model) and the extent to which the model actually describes individual preferences (i.e., the descriptive validity of the model). Empirical evidence for the descriptive validity of the QALY assumptions is fairly scarce. In a review of the literature available to that date, Loomes and McKenzie' drew rather negative conclusions, but examined only whether the conditions hold exactly. They did not allow for imprecision in respondents' preferences. Miyamoto and Eraker[3] tested utility independence of quantity of life from quality of life. Their results support this condition. With respect to the tradeoff between quality of life and quantity of life, several authors have observed that increasing proportional tradeoff (i.e., an individual is willing to sacrifice relatively more remaining life years when the number of remaining life years is greater) is more consistent with individual preferences than is constant proportional tradeoff."* Moreover, in some studies it was observed that for short time durations individuals were not willing to trade any life years for an improvement in quality of life.[3,4] Apparently, individual preferences are lexicographic for short time durations: individual choices are fully determined by the number of life years remaining.

More evidence is available with respect to risk neutrality. Studies that directly tested risk neutrality on life years by assessing utility functions over life years typically reject risk neutrality.[5-8] The one exception is the study by Miyamoto and Eraker[9] that found risk neutrality to hold for the hypothetical "average respondent." Even in that study, however, there were few respondents whose preferences actually satisfied risk neutrality on life years.

A disadvantage of the characterization of the QALY model by Pliskin et al. is that it applies only to health profiles of a constant quality. More general characterizations of the QALY model exist that allow quality of life to vary over time.""" An advantage of the characterization by Pliskin et al. is that utility independence and constant proportional tradeoff are directly related to two commonly used methods of estimating quality weights for health states: the standard gamble and the time tradeoff, respectively. These relationships allow straightforward tests of utility independence and constant proportional tradeoff in an experimental, design. We report the results of an experiment aimed at testing constant proportional tradeoff and utility independence of quality of life from quantity of life. To the best of our knowledge, the latter condition has not been tested before. Miyamoto and Eraker[3] tested the converse: whether quantity of life is utility-independent from quality of life. The importance of testing whether constant proportional tradeoff and utility independence hold simultaneously follows from the rejection of risk neutrality on life years in various studies. The results of these studies challenge the descriptive validity of the risk-neutral QALY model. The question then emerges whether the more general risk-adjusted QALY model performs better in describing individual preferences. As explained above, the risk-adjusted QALY model is characterized by 'constant proportional tradeoff and utility independence. Testing these two conditions provides insight in the descriptive validity of the risk-adjusted QALY model.

Testing constant proportional tradeoff and utility independence separately is also interesting in its own right. Constant proportional tradeoff ensures that time tradeoff weights are independent of the time horizon used in the assessment. Utility independence ensures that standard-gamble weights are independent of the time horizon used in the assessment. Utility independence also allows the estimation of another generalization of the risk-neutral QALY model, in which risk neutrality and constant proportional tradeoff are re1axed.[11]

The structure of the paper is as follows. In the next section we explain in more detail the theory of QALYs. Then the experimental methods and results are discussed. Finally, the results and the implications of our findings for the use of QALYs in medical decision making are discussed.

Theoretical Analysis of QALYs

We confine ourselves to an analysis of preferences over health profiles of constant quality. Let (Q, T) denote a health profile consisting of T years in quality of life level Q. Let a typical gamble over quality of life and quantity of life in which health profile (Q_i, T_i) occurs with probability p_i be denoted by $[p_1, (Q_1, T_1); p_2, (Q_2, T_2); \ldots; p_n, (Q_n, T_n)]$. All quality-of-life levels are assumed to be more attractive than death. Further, all $T_i \geq 0$, all $p_i \geq 0$, and $\Sigma p_i = 1$. We assume that an individual preference relation over gambles involving quality of life and quantity of life satisfies the axioms of von Neumann-Morgenstern (vNM) expected-utility theory." Then a real-valued, cardinal, utility function $U(Q, T)$ exists, the expected value of which represents individual preferences over gambles involving quality of life and quantity of life.

Pliskin et al. have derived that QALYs are a valid

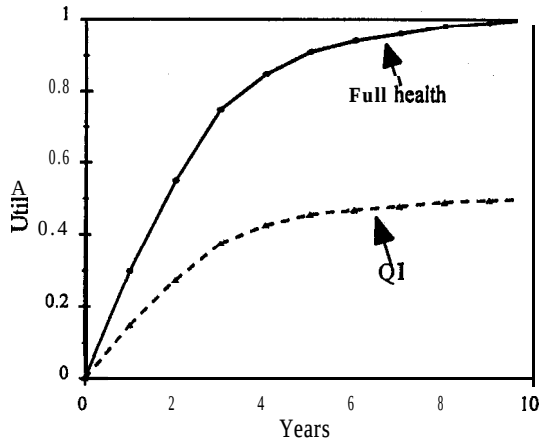

FIGURE 1. The utility function for life years under mutual utility independence.

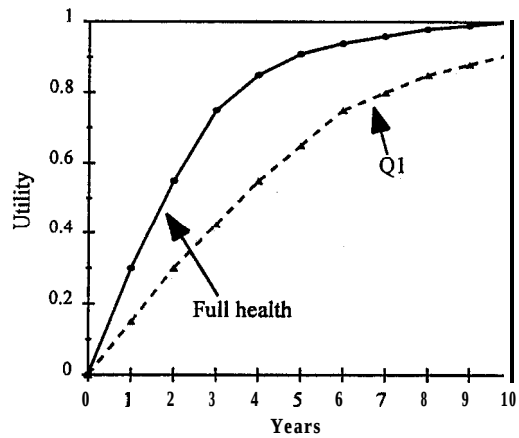

FIGURE 2. The utility function for life years under constant proportional tradeoff.

vNM utility function if in addition to the vNM axioms three other conditions are imposed on the individual preference relation: (mutual1 utility independence, constant proportional tradeoff, and risk neutrality on life years. We consider each in turn.

UTILITY INDEPENDENCE

When we fix one of the two attributes in the utility function over health profiles at a particular value, utility independence imposes that preferences with respect to gambles over the other attribute do not depend on the particular value chosen. Formally, utility independence implies that $[p_1, (Q_1, T_1); p_2, (Q_2, T_1); \ldots; p_n, (Q_n, T_1)]$ is preferred to $[r_1, (Q_1, T_1); r_2, (Q_2, T_1); \ldots; r_n, (Q_n, T_1)]$ if and only if $[p_1, (Q_1, T_2); p_2, (Q_2, T_2); \ldots; p_n, (Q_n, T_2)]$ is preferred to $[r_1, (Q_1, T_2); r_2, (Q_2, T_2); \ldots; r_n, (Q_n, T_2)]$ for all T_1, T_2. A similar expression holds when Q is held fixed and T varies. Denote by $W(Q)$ a utility function over quality of life and by $V(T)$ a utility function over quantity of life. Keeney and Raiffa[13] have shown that utility independence implies that $U(Q, T)$ is either multiplicative, i.e., $W(Q) \cdot V(T)$, or additive, i.e., $W(Q) + V(T)$.* The additive model depends on a condition Pliskin et al. refer to as "marginality." Pliskin et al. and Miyamoto and Eraker[3,9] provide arguments why the additive model is not realistic in the medical context. The additive model can be excluded by adding an entirely plausible condition to the model: that for a time duration of zero life years the individual is indifferent between all quality-of-life levels.† If the additive model is discarded, utility independence can be shown to imply: $U(Q_1, T_1)/U(Q_2, T_1) = U(Q_1, T_2)/U(Q_2, T_2)$.‡ If we plot $U(Q, T)$ against T, holding quality of life fixed, then utility independence guarantees that the shape of $U(Q, T)$ is the same regardless of the level at which quality of life is held fixed. This is illustrated in figure 1 for life durations up to ten years, where, for convenience, full health is selected as quality-of-life level Q_2. If utility independence holds, then for all health states the fraction of the utility of full health is independent of the time horizon. In figure 1, for instance, the utility of health state Q_1 is 0.5 of the utility of full health for all time horizons.

'Utility independence facilitates the determination of standard-gamble quality weights. The standard gamble determines the quality weight of a health state by comparing a specific number of years in this health state with a gamble with a probability (p) of the same number of years in full health and a complementary probability $(1 - p)$ of immediate death. The probability of full health (p) is varied until the individual is indifferent between the alternatives. Suppose W(Q) is scaled such that W(full health1 = 1 and $W(\text{death}) = 0$. The quality weight of the health state is then set equal to p^*, where p^* is the probability for which the individual is indifferent. Thus, the standard gamble measures the utility of a health state as the fraction of the utility of full health. By utility independence, this fraction does not depend on the number of years used in the measurement. The only restriction is that the number of life years be the same for the certain health state and for full health. In figure 1, the standard-gamble quality weight for Q_1 is equal to 0.5 regardless of the time horizon used in the assessment.

*As one of the referees reminded us, this holds only when V and W are rescaled in line with U. For more details see Keeney and Raiffa[13] (pp. 289-91).

†A proof of this result has been provided elsewhere.[14]

‡$U(Q_1, T_1)/U(Q_2, T_1) = [W(Q_1) \cdot V(T_1)]/[W(Q_2) \cdot V(T_1)] = W(Q_1)/W(Q_2) = [W(Q_1) \cdot V(T_2)]/[W(Q_2) \cdot V(T_2)] = U(Q_1, T_2)/U(Q_2, T_2)$. A similar argument shows that $U(Q_1, T_1)/U(Q_1, T_2) = U(Q_2, T_1)/U(Q_2, T_2)$.

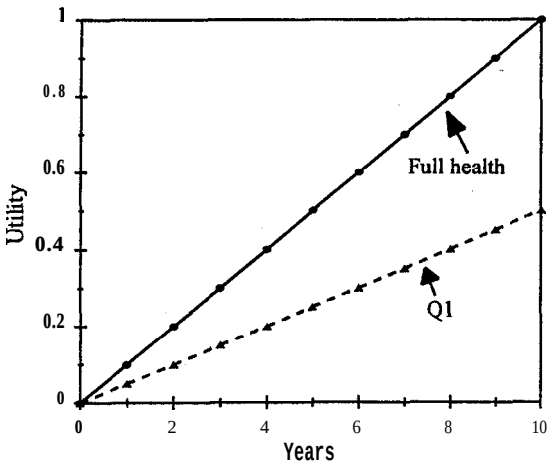

FIGURE 3. The utility function for life pars under risk neutrality.

CONSTANT PROPORTIONAL TRADEOFF

Constant proportional tradeoff imposes that if an individual is indifferent to a choice between T years in health state Q_1 and αT years $(0 \leq \alpha \leq 1)$ in a more attractive health state Q_2, then this individual should also be indifferent between βT years $(\beta \geq 0)$ in Q_1 and $\beta\alpha T$ years in Q_2. The proportion "years in Q_1 divided by years in Q_2" is constant by the condition of constant proportional tradeoff (in the above choice situation this proportion is equal to $1/\alpha$). Constant proportional tradeoff is illustrated in figure 2, where the individual is willing to sacrifice 50% of his remaining life span in Q_1 to improve his health to Q_2, which is set equal to full health for convenience. By constant proportional tradeoff, this proportion holds for any time horizon. Thus, as can be seen from figure 2, the individual is indifferent between six years in Q_1 and three years in full health, but also between four years in Q_1 and two years in full health.

Constant proportional tradeoff facilitates the assessment of time-tradeoff quality weights. The time tradeoff determines the quality weight of a health state by comparing T years in the health state with X years in full health. The number of years in full health (X) is varied until the individual is indifferent between the alternatives. The quality weight of the health state is then set equal to X/T. The time tradeoff thus measures quality weights as the equivalent fractions of healthy years. By the assumption of constant proportional tradeoff, this fraction will be independent of the time horizon used in the assessment.

Denote the time tradeoff weight by $W_1(Q)$ and the standard-gamble weight by $W_2(Q)$. Pliskin et al. have shown that if both utility independence and constant proportional tradeoff hold, individual preferences can be represented by a risk-adjusted QALY model: $[T \cdot W_1(Q)]^r = T^r W_2(Q)$. The parameter r in this equation reflects the individual's attitude to risk with respect to survival duration. It is clear from the equation that both standard gamble and time tradeoff can be used to determine quality weights for the calculation of the number of risk-adjusted QALYs. However, in general they will not give identical quality weights. Time-tradeoff weights have to be adjusted by the risk parameter to arrive at standard-gamble weights.

RISK NEUTRALITY ON LIFE YEARS

The only situation in which the standard gamble and the time tradeoff will, at least in theory, elicit identical quality weights is the situation in which $r = 1$, i.e., the situation in which the individual is risk-neutral with respect to life years. This situation characterizes the QALY measure most frequently used in cost-utility analysis. Risk neutrality with respect to life years implies a utility function for life years that is linear in life years. Figure 3 illustrates a risk-neutral utility functions both for full health and for a health state Q_1. In figure 3, both the time-tradeoff weight and the standard-gamble weight are equal to 0.5.

Methods

RESPONDENTS

The respondents were 80 students at the Stockholm School of Economics and 92 students at Erasmus University Rotterdam. All were undergraduates recruited from courses in economics, statistics, and health policy. Each was paid approximately $15 for participating. The experiment was carried out in 17 sessions lasting approximately one hour with, on average, ten respondents per session. The procedure in each session was to explain a specific task to the respondents, obtain their responses to this task, and then move on to the next task. A "master" version of the experiment was designed in English. This "master" version was subsequently translated into Swedish and Dutch. Before drafting the final version, we tested the questionnaire extensively both in Stockholm and in Rotterdam, using faculty staff members as respondents.

HEALTH STATES

The health states included in the questionnaire were taken from the Maastricht Utility Measurement Questionnaire, an adaptation of the McMaster Health Utility Index,[15,16] and correspond to commonly occurring types of back pain and rheuma-

tism. Each health state consisted of four dimensions: general daily activities, self-care, leisure activities, and pain. The health states were described on a set of cards that were handed out to respondents at the beginning of each session. Health states B and D, relevant for the analysis reported in this paper, are presented in table 1; state B is clearly more attractive than D.

QUESTIONNAIRE

The first two tasks consisted of the ranking and placing on a rating scale calibrated from 100 (full health) to 0 (immediate death) of six health states. There were two versions of the questionnaire, each administered in different sessions. For reasons not related to the present study, two of the six health states varied per version. The two health states varied in version 1 were less attractive than the two varied in version 2. For every respondent, however, health states B and D were included. The possibility exists that the inclusion of different health states in the two versions affected the results of our experiment. We return to this possibility after the description of the experimental tests.

Constant proportional tradeoff was subsequently tested. Time-tradeoff quality weights for health states B and D were determined. The respondents were instructed to indicate first the values of X, the number of healthy years, for which they definitely preferred to be in health states B and D, respectively, then the values of X for which they definitely preferred to be in full health, and finally those values of X for which they found it hard to choose between the alternatives. The respondents were explicitly told that all profiles would result in death after the indicated number of years. The general introduction to the time-tradeoff questions can be found in appendix A. It was pointed out to the respondents both in the text and in the oral explanation of the task that they could indicate a range of values for X for which they found it hard to choose between the alternatives. The response strategy we suggested to the respondents is likely to lead to an interval of indifference values. When respondents first mark all the values for which they have clear preferences for one of the alternatives, they end up with a range of values for which they are not certain which alternative to prefer. We told the respondents to mark these values with the symbol for indifference. This format was adopted to allow for the fact that respondents are likely to have imprecise preferences." Respondents are unfamiliar both with the health states to be assessed and with the idea of trading off life years. As a result, their preferences may be imprecise. In our format, we attempted to take this imprecision into account. Where respondents indicated ranges of values for which they could not choose between the options, we interpreted these ranges of values as their personal confidence intervals (PCIS).

Both versions of the questionnaire contained three time-tradeoff questions. Version 1 presented a time-tradeoff question for ten years in health state D (DlO), followed by a question for 30 years in health state D (D30) and a question for 30 years in health state B (B30). The sequence in version 2 was Bl0, B30, D30. This setup allowed a test of constant proportional tradeoff at the individual level. For version 1, we compared the responses to questions Dl0 and D30. For version 2, we compared the responses to questions Bl0 and B30.

In section 4, utility independence of quality of life from quantity of life was tested.5 Standard-gamble quality weights for health states B and D were determined. Again, the respondents were explicitly informed that all health profiles would be followed by death. Probability elicitation was by means of a line of values for the probability of successful treatment (full health). Next to this line a line was drawn with the complementary probability of failure of treatment (immediate death). This display was chosen in an attempt to control for a potential framing bias: displaying only the probability of successful treatment might induce an individual to focus on successful treatment, not sufficiently taking into account the probability of failure of treatment. Psychological evidence of the influence of reference effects on choice is abundant." As with the time

Table 1 • Health States B and D

B
• Able to perform all tasks at home and/or at work, albeit with some difficulties
• Able to perform all self-care activities (eating, washing, dressing) without help
• Unable to participate in certain types of leisure activities
• Often light to moderate pain and/or other complaints
D
• Unable to perform some tasks at home and/or at work
• Able to perform all self-care activities (eating, washing, dressing), albeit with some difficulties
• Unable to participate in many types of leisure activities
• Often moderate to severe pain and/or other complaints

§In the remainder of this paper we for convenience use the term "utility independence" when we mean utility independence of quality of life from quantity of life.

tradeoff question, an attempt was made to take imprecision of preferences into account. First, the respondents were asked to indicate those values of p, **the** probability of successful treatment, for which they definitely preferred the certain option, then those values of p for which they definitely preferred the treatment option (gamble), and finally those values of p for which they found it hard to choose between the options. The general explanation of the standard-gamble questions can be found in appendix B. Again, it was pointed out to the respondents both in the description of the task and in the oral explanation that they were allowed to indicate ranges of values of p for which they found it hard to choose between the options. These ranges of values were then interpreted as their PCIs for p.

Section 4 was also administered in two versions. In version 1, the sequence of the questions was Dl0, D30, B30; in version 2, the sequence of the questions was Bl0, B30, D30. Utility independence was tested by comparing the responses to questions Dl0 and D30 in version 1 and to Bl0 and B30 in version 2.

Personal confidence intervals should not be confused with statistical confidence intervals, but the idea behind them is somewhat similar. A PCI indicates a range of values for which an individual expresses indifference and which cannot be distinguished from the "true" indifference value. Our interpretation of PCIs is similar to the interpretation of statistical confidence intervals: if PCIs of two values overlapped, we interpreted the difference between the values as not significant in terms of the individual's PCI. When the PCIs did not overlap, the difference was interpreted to be significant.

In the Results section, we present the responses both with and without adjustment for imprecision of preferences. For those respondents who indicated PCIs we used this interval to adjust for imprecision of preference. For those respondents who did not indicate a PCI, we constructed an artificial PCI by adjusting for the median imprecision of preference, computed from the responses of those respondents who had indicated PCIs. For comparison we also present the results when no artificial PCIs were constructed. These results interpret the responses of those respondents who did not indicate a PCI as being precise. It should be emphasized here that we find it hard to believe that these respondents were indeed certain of their responses, given the unfamiliarity of the health states and of the tasks they were requested to perform. We return to this in the final section of the paper.

Two types of biases may have affected the data. First, asking the questions for ten and 30 years in fixed sequence may have caused responses to be anchored. For example, respondents may adopt a proportional heuristic in answering the time-tradeoff questions. That is, they simply state a fixed percentage of the remaining lifetime, even though their preference relation does not actually satisfy constant proportional tradeoff: In answering the standard-gamble questions, respondents may likewise simply give the same indifference probability to minimize cognitive effort. To control for the possibility of an anchoring strategy, we included between-subjects tests of constant proportional tradeoff and utility independence, which are not affected by anchoring: the mean weights for B30 in version 1 were compared with the mean weights for Bl0 in version 2. Similarly; the mean weights for D30 in version 2 were compared with the mean weights for Dl0 in version 1. Some inferences with respect to the effect of anchoring can be drawn from a comparison of the responses to B30 and D30 between the two versions. For example, given increasing proportional tradeoff, one would expect the time-tradeoff weight for D30 to be higher in version 1, in which Dl0 was also included. If the individual preference relation satisfies increasing proportional tradeoff, Dl0 will in general be higher than D30. However, anchoring has the effect of making Dl0 and D30 equal. Therefore, if Dl0 is asked first, given increasing proportional tradeoff, anchoring will induce an upward bias on **D30.¶** Obviously, the upward bias of anchoring will occur only with version 1, because in version 2 Dl0 was not included in the questionnaire. Therefore, we expect D30 to be higher in version 1 if anchoring is a problem. By a similar line of reasoning, B30 can be expected to be higher with version 2, given increasing proportional tradeoff. This test assumes a random distribution of preferences in the two samples. We have no reason to believe that this assumption does not hold. Respondents were allocated randomly to the versions, and we have no indication that significant bias was introduced by the allocation process.

The second bias may have been introduced by the fact that the versions differed in the six health states that were evaluated. As explained above, version 1 contained more severe health states than version 2. This may have made health states B and D appear more attractive, and may thus have resulted in higher weights for health states B and D, in version 1. Two points are worth making with respect to this possible bias. First, it will affect only our between-subjects tests. Within subjects obviously the same version was used and we can still compare the answers. Second, if this bias indeed affects our data, we would expect it to be stronger for health state B

¶Obviously, under decreasing proportional tradeoff and anchoring the opposite pattern holds: D30 will be lower in version 1.

than for health state D because health state D was still the worst health state in both versions. This was reflected by the ranking exercise: all but two (version 2) respondents ranked health state D as the worst health state. Analysis of the rating-scale valuations confirmed this expectation. The rating-scale valuations for health state B differed significantly across versions, but the difference between the rating scale valuations was not significant for health state D. Therefore, if the two versions differed significantly in the weights for health state B, but to a smaller extent in the weights for health state D, we interpret this response pattern as an indication that the inclusion of different health states has produced a bias in the responses.

Statistical Analysis

Mean values within samples were compared by means of two-tailed paired t-tests. The paired t-test assumes normality of differences, but is fairly robust. To be on the safe side, when normality was rejected by a Kolmogorov-Smirnov test, we tested for equality of means by the nonparametric Wilcoxon matched-pairs signed-rank sum test. Mean values between samples were compared by two-tailed independent-samples t-tests. The independent-samples t-test is robust for non-normality if the hypothesis of equal variances in the two samples cannot be rejected. We therefore first tested equality of variances by means of an F-test. If equality of variances was rejected, the nonparametric Mann-Whitney test was used to analyze the data.

Given the size of our sample, at a significance level of 5%, the paired t-test is able to detect a difference of 0.25 times the standard deviation with a power of over 90%. Given that standard deviations for time-tradeoff and standard-gamble quality weights reported in the literature rarely exceed 0.2,‖ the probability of detecting a true difference of 0.05 by the paired t-test is higher than 90%. The power of the independent-samples t-test to detect a difference of 0.25 times the standard deviation at a significance level of 5% is 45%. The power to detect a difference of 0.5 times the standard deviation is higher than 90%.

Hypotheses with respect to proportions were tested by calculating χ^2 values from the resulting 2 X 2 contingency tables, which were compared with *1* degree of freedom. Continuity corrections were made both in the case where proportions come from the same population and in the case where proportions come from different populations. The method used to test hypotheses with respect to proportions uses the continuous normal distribution as an approximation of the discrete binomial distribution. The normal distribution corresponds better to the binomial distribution when a correction is made to the observed frequency to allow for the fact that variables can take only integer values.[19]

To examine whether a systematic relationship exists between satisfying constant proportional tradeoff and satisfying utility independence, we used a binary choice model. We took the O-l variable "satisfying constant proportional tradeoff yes/no" as the variable to be explained. This variable takes the value 0 if a respondent does not satisfy constant proportional tradeoff and 1 if a respondent does satisfy constant proportional tradeoff. The O-l variable "satisfying utility independence yes/no" was taken as the explanatory variable. This variable takes the value 0 if a respondent does not satisfy utility independence and 1 if a respondent does satisfy utility independence. For this analysis, we could choose between a probit model, in which the error terms are distributed according to the standard normal distribution, and a logit model, in which the error terms are distributed according to the logistic distribution. Because we estimated a univariate dichotomous model, it is hard to distinguish between the two methods.[“] However, the logistic distribution has slightly heavier tails, and, because we could not exclude the possibility that responses to the constant proportional tradeoff and utility independence questions would be concentrated in the tails, we decided to use the logit model. Model performance was assessed by the likelihood-ratio test.

Results

We present the results for the two samples combined. Separate results for the Swedish and Dutch samples are not reported, because the results were not significantly different. Separate results for the two samples are available from the authors upon request.

ANALYSIS OF INDIVIDUAL RESPONSES

We comment on the aggregated data only, because no significant difference was found between the two versions in the individual analyses.

Constant proportional tradeoff The first three lines of table 2 show the results of the test of constant proportional tradeoff on the basis of the individual data. Individuals whose choices exactly satisfy constant proportional tradeoff are in category C (22.7% of respondents). The proportion of respondents in category C is slightly distorted, because three respondents were not willing to trade any life

‖Most standard deviations in our study were also less than 0.2.

Table 2 • Numbers of Time-tradeoff (TTO) and Standard-gamble (SG) Responses per Category*

	Version(s)	*n*	A	B	C	D	E
Time tradeoff	1	87	5	31	18	30	5
Time tradeoff	2	85	5	27	23	28	2
Time tradeoff	1 and 2	172	10	58	39	58	7
Standard gamble	1	87	18	38	8	17	10
Standard gamble	2	85	12	44	15	10	4
Standard gamble	1 and 2	172	28	80	23	27	14

*The categories:
A = weight 10 years > weight 30 years and no overlap of personal confidence intervals (PCIs).
B = weight 10 years > weight 30 years, but overlap of PCIs.
C = weight 10 years = weight 30 years.
D = weight 10 years < weight 30 years, but overlap of PCIs.
E = weight 10 years < weight 30 years and no overlap of PCIs.

Table 3 • Numbers of Respondents Satisfying Constant Proportional Tradeoff (CPT) and Utility Independence (UI) Simultaneously, Unadjusted and with Adjustment for Imprecision of Preference

	n	CPT + UI	CPT_{adj} + UI_{adj}
Version 1	87	3	54
Version 2	85	7	84
Versions 1 and 2	172	10	118

years at all for an improvement in health. Contrary to previous studies,[4] table 2 shows no indication that increasing proportional tradeoff is a more common response pattern than decreasing proportional tradeoff.

These results do not take into account that respondents' preferences are probably somewhat imprecise. For those respondents who indicated personal confidence intervals, we used these intervals to examine overlap. However, such confidence intervals were given for only 92 responses.** The artificial confidence interval estimated on the basis of the median imprecision of preference?? resulted in a personal confidence interval of [TTO – 0.075; TTO + 0.0751.

The respondents in categories B and D had overlapping personal confidence intervals. If the difference between the TTO valuations is interpreted as not significant, these respondents are counted as satisfying constant proportional tradeoff. The respondents in categories A and E had non-overlapping personal confidence intervals. Their choices violate constant proportional tradeoff even after adjustment for imprecision of preference.

Table 2 shows that the majority of time-tradeoff responses satisfied constant proportional tradeoff with imprecision adjustment (90.1%). The differences between the proportions satisfying increasing proportional tradeoff and decreasing proportional tradeoff are not significant in both versions. If responses are only partially adjusted for imprecision of preference (i.e., no artificial personal confidence intervals are constructed) 37.0% of the respondents satisfy constant proportional tradeoff.

Utility independence. The last three lines of table 2 show that the overall proportion of respondents who exactly satisfy utility independence, the respondents in category C, is 13.4%. The proportion of respondents in categories A and B is significantly higher than the proportion of respondents in D and E [$\chi^2(1) > 10.8$; p C 0.00 in both versions]. This indicates that the utilities of health states B and D as fractions of the utility of full health decrease with the time horizon used in the assessment. Utility independence predicts that these fractions should be constant.

The artificial personal confidence interval constructed for those respondents who did not indicate a personal confidence interval is equal to: [SG – **0.05; SG + 0.0513** Table 2 shows that after adjustment for imprecision of preference majorities of the respondents in the various groups satisfy utility independence. However, the proportion of respondents who satisfy utility independence (75.6%) is lower than the proportion of respondents who satisfy constant proportional tradeoff. If only a partial adjustment is made for imprecision of preference, the proportion of respondents in categories B, C, and D increases only slightly, to 18.0%.

Constant proportional tradeoff and utility independence. Pliskin et al.' have derived that constant proportional tradeoff and utility independence guarantee that individual preferences over lotteries on chronic health profiles can be represented by the risk-adjusted QALY model. In table 3, the column CPT + UI shows the number of respondents who satisfy both constant proportional tradeoff and utility independence when no adjustment is made for imprecision of preference. The column CPT_{adj} + UI_{adj} shows the number of respondents who satisfy both

**The fact that this number is equal to the number of respondents in the Dutch survey is pure coincidence.

††Adjusting for the mean imprecision of preference resulted in slightly larger personal confidence intervals: [TTO – 0.09; TTO + 0.091. However, using the mean imprecision of preference to construct artificial personal confidence intervals hardly affected the results: both in version 1 and in version 2 one individual no longer violated constant proportional tradeoff with imprecision adjustment.

‡‡The mean imprecision of preference was approximately the same: 0.052. Using personal confidence intervals estimated on the basis of the mean imprecision of preference did not affect the results.

constant proportional tradeoff and utility independence when responses are adjusted for imprecision of preference. Table 3 shows that, unadjusted for imprecision of preference, only 5.8% of the respondents satisfy the two conditions simultaneously. After adjustment for imprecision of preference@ 68.6% of the respondents satisfy the risk-adjusted QALY model. Partial adjustment for imprecision of preference only marginally increases the proportion of respondents who satisfy both constant proportional tradeoff and utility independence: the proportion rises from 5.8% to 10.5%.

Table 3 does not show whether a systematic relationship exists between constant proportional tradeoff and utility independence. Is a respondent who satisfies utility independence also more likely to satisfy constant proportional tradeoff? We estimated logistic regressions to examine the existence of such a systematic relationship. Denote the probability that a respondent has a value i on the constant proportional tradeoff variable given that this respondent has a value j on the utility independence variable by P(CPT = i|UI = j). For example, the probability that a respondent satisfies constant proportional tradeoff given that he or she satisfies utility independence is denoted by P(CPT = 1| UI = 1).

Table 4 lists the results of the estimation procedure. Results are reported only for the situation where no adjustment for imprecision of preference was made. Information about whether or not a respondent satisfies utility independence does not improve the model significantly if imprecision adjustment is applied. This is caused by the unequal distribution of observations over cells, which in turn is a consequence of the fact that with imprecision adjustment most respondents satisfy both constant proportional tradeoff and utility independence. Table 4 shows that in every situation respondents who satisfy utility independence are more likely to satisfy constant proportional tradeoff. The contribution of the model is significant in all but one case.

Table 4 • Results of the Logistic Regression Estimation without Adjustment for Imprecision of Preference*

	P(CPT = 1|UI = 0)	P(CPT = 1|UI = 1)	Model χ^2
Version 1	18.5%	37.5%	1.81 (NS)
Version 2	22.9%	48.7%	3.27 (p = 0.0708)
Versions 1 and 2	19.5%	43.5%	5.78 (p = 0.0182)

* P(CPT = 1|UI = 0) denotes the probability that a respondent satisfies constant proportional tradeoff (CPT) given that he or she does not satisfy utility independence (UI). The model χ^2 has been calculated by the likelihood-ratio test.

Table 5 • Mean Time-tradeoff and Standard-gamble Weights (Standard Errors in Parentheses)

	Health State*	*n*	Time Tradeoff	Standard Gamble
Version 1	D, 10 years	87	0.5901 (.0194)	0.7017 (.0249)
Version 1	D, 30 years	87	0.5893 (.0201)	0.8784 (.0256)
Version 1	B, 30 years	87	0.8045 (.0155)	0.8851 (.0181)
Version 2	B, 10 years	85	0.7947 (.0169)	0.8972 (.0145)
Version 2	B, 30 years	85	0.7841 (.0179)	0.8507 (.0191)
Version 2	D, 30 years	85	0.5684 (.0273)	0.8597 (.0282)

*See table 1.

GROUP ANALYSIS

Constant proportional tradeoff Column four of table 5 shows the mean values for time tradeoff. Within groups, constant proportional tradeoff predicts equality between Dl0 and D30 in version 1 and between Bl0 and B30 in version 2. The difference in the weights between Dl0 and D30 and Bl0 and B30 is in both cases not significant, and the null hypothesis of no difference-cannot be rejected. The aggregate pattern is consistent with constant proportional tradeoff. We observe no indication of anchoring: the null hypothesis that the weights for B30 and D30 are equal in the two versions, cannot be rejected.

The between-groups test also provides support for constant proportional tradeoff. We compared the version 1 responses to Dl0 with the version 2 responses to D30 and the version 1 responses to B30 with the version 2 responses to Bl0. In both comparisons the null hypothesis of equal weights cannot be rejected.

Table 5 does not indicate that a bias has been introduced by the fact that different health states were included in the two versions since the weights for B30 and D30 are not significantly different between the versions. Recall from the argument outlined in the Methods section that if a bias had been introduced, we would expect the weights for both B30 and D30 to be higher with version 1. Moreover, we would expect the difference to be more pronounced for B30.

§§For respondents who did not indicate a personal confidence interval, we used the same artificial personal confidence intervals as before to adjust for imprecision of preference: [TTO − 0.075; TTO + 0.0751 and [SG − 0.05; SG + 0.051.

Utility independence. The last column of table 5 shows the standard-gamble weights. The results of the within-groups analysis do not provide support for utility independence. Utility independence predicts equality between Dl0 and D30 in version 1 and between Bl0 and B30 in version 2. However, the weight for 10 years is significantly higher than the weight for 30 years, both for health state B ($p < 0.001$) and for health state D ($p < 0.05$). There is no indication of anchoring: the weights for B30 and D30 do not differ significantly between the versions.

The between-groups test of utility independence also suggests violation of utility independence. Both Dl0 and Bl0 are higher than D30 and B30, respectively, which confirms the pattern observed within groups. However, the differences are not significant. This may be due to the lower power of the independent-samples t-test.?!

There is no indication that a bias has been introduced by the difference in included health states in the two versions. The weights for B30 and D30 are higher in version 1, but the difference is not significant and, contrary to expectation; the difference is more pronounced for health state D than for health state B.

Discussion

We have experimentally tested two of the preference conditions that underlie the QALY model in the derivation by Pliskin et al.‘: utility independence and constant proportional tradeoff. The results suggest that constant proportional tradeoff is a condition that describes individual preferences reasonably well. The group analysis revealed that both within- and between-groups constant proportional tradeoffs could not be rejected. The analysis of the individual responses showed that 22.8% of the respondents satisfied constant proportional tradeoff without adjustment for imprecision of preference. Deviations from constant proportional tradeoff are not systematic: increasing proportional tradeoff and decreasing proportional tradeoff were observed with approximately equal frequencies. After adjustment for imprecision of preference, which is likely to have occurred given the respondents' relative unfamiliarity both with the health states and with the methods of utility measurement, the proportion of respondents whose choices satisfy constant proportional tradeoff increased to 90.1%. With partial adjustment for imprecision of preference, this proportion increases up to 37.0%.

Our results provide less support for utility independence. Within groups, the fraction of the utility of full health decreased rather than staying constant as predicted by utility independence. The between-groups analysis also suggested violation of utility independence. However, in this case the violations were not statistically significant, and we could not reject utility independence, probably due to the lower power of the independent-samples t-test. The analysis of the individual responses showed that 13.4% of the respondents satisfied utility independence without adjustment for imprecision of preference. After adjustment for imprecision of preference, this proportion increased to 75.6%. With partial adjustment for imprecision of preference, this proportion increased only marginally, to 18.0%.

Pliskin et al.' have derived that imposing both constant proportional tradeoff and (mutual) utility independence ensures that individual preferences over lotteries over chronic health profiles can be represented by a risk-adjusted QALY model. In our study, 5.8% of the respondents satisfied both constant proportional tradeoff and utility independence (of quality of life from quantity of life) when no adjustment was made for imprecision of preference. When adjustment was made for imprecision of preference, this proportion increased to 68.6%. With partial adjustment, the proportion increased to 10.5%.

Adjustment for imprecision of preference turns out to have an important influence on the results of the individual analysis. It should be remembered that for those respondents who did not indicate a personal confidence interval, a personal confidence interval had to be estimated. Estimation of a personal confidence interval is necessarily an arbitrary exercise. However, in our opinion it is unlikely that the actual personal confidence intervals are wider than the estimated personal confidence intervals for these respondents. The fact that exact responses were given even though the possibility of indicating personal confidence intervals had been pointed out to the respondents suggests that these respondents may have had reasonably precise preferences.

We therefore believe that our estimates with adjustment for imprecision of preference should be considered to be maximum estimates. The partially adjusted results (in which no artificial personal confidence intervals have been constructed and only reported personal confidence intervals have been used) should in our interpretation be considered minimum estimates.

The fact that a large proportion of the respondents did not indicate a personal confidence interval

¶¶This may appear somewhat surprising because D30 in version 2 is lower than D30 in version 1, for example, and their standard errors are approximately equal. However, within versions the paired t-test is used, in which correlation between Dl0 and D30 is taken into account. Between-versions independence of valuations is assumed. The independence assumption results in larger standard errors and therefore lower t-values.

even though we encouraged them to do so is somewhat surprising. This may have been due to the fact that we did not require the respondents to indicate personal confidence intervals, but only included this as an option. Indicating an interval is not necessarily easier than indicating one value. Indication of an interval requires careful thinking about upper and lower bounds. The respondents who did not state an interval may have found the cognitive effort to provide just one value less demanding. Due to the uncertainty about the personal confidence intervals for the respondents who did not indicate an interval, our results about the proportion of respondents satisfying the preference conditions after adjustment for imprecision of preference need to be interpreted with great care. To get more definite results, it may be necessary in future research to require that respondents indicate personal confidence intervals.

These results indicate that constant proportional tradeoff holds approximately. Constant proportional tradeoff has clear implications only if quality of life and quantity of life are utility independent. However, the evidence is much weaker for utility independence. More research aimed both at testing constant proportional tradeoff and utility independence in different experimental settings and at developing alternative utility models in health remains necessary.

The authors are grateful to Eddy ban Doorslaer, Peter Wakker, John Miyamoto, the editor, and two anonymous referees for their comments on previous drafts; to Maureen Rutten-van Mölken for helpful suggestions with respect to the selection of the health states; and to Jaco van Rijn for assistance in running the experimental sessions.

References

1. Pliskin JS, Shepard DS, Weinstein MC. Utility functions for life years and health status. Oper Res. 1980;28:206-24.
2. Loomes G, McKenzie L. The use of QALYs in health care decision making. Soc Sci Med. 1989;28:299-308.
3. Miyamoto JM, Eraker SA. A multiplicative model of the utility of survival duration and health quality. J Exp Psycho1 Gen. 1988;117:3-20.
4. McNeil BJ, Weichselbaum R, Pauker SG. Speech and survival: tradeoffs between quality and quantity of life in laryngeal cancer. N Engl J Med. 1981;305:982-7.
5. Stiggelbout AM, Kiebert GM, Kievit J, Leer JWH, Stoter G, de Haes JCJM. Utility assessment in cancer patients: adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. Med Decis Making. 1994;14:82-90.
6. McNeil BJ, Weichselbaum R, Pauker SG. Fallacy of the five-year survival in lung cancer. N Engl J Med. 1978;299:1397-401.
7. Verhoef LCG, de Haan AFJ, van Daal WAJ. Risk attitude in gambles with years of life: 'empirical support for prospect theory. Med Decis Making. 1994;14:194-200.
8. Maas A, Wakker PP. Additive conjoint measurement for multiattribute utility. J Math Psychol. l994;38:86-101.
9. Miyamoto JM, Eraker SA. Parameter estimates for a QALY utility model. Med Decis Making. 1985;5:191-213.
10. Broome J. QALYs. J Public Econ. 1993;50:149-67.
11. Bleichrodt H. QALYs and HYEs: under what conditions are they equivalent? J Health Econ. 1995;14:17-37.
12. von Neumann J, Morgenstern 0. Theory of Games and Economic Behavior. Princeton, NJ: Princeton University Press, 1947.
13. Keeney R, Raiffa H. Decisions with multiple objectives: preferences and value tradeoffs. New York: Wiley, 1976.
14. Bleichrodt H, Wakker PP, Johannesson M. Characterizing QALYs by means of risk neutrality. Working paper, Erasmus University Rotterdam, Rotterdam, The Netherlands, 1995.
15. Bakker C, Rut-ten-van Mölken M, van Doorslaer E, Bennett K, van der Linden Sj. Feasibility of utility assessment by rating scale and standard gamble in ankylosing spondylitis or fibromyalgia. J Rheumatol. 1995;22:1536-43.
16. Rutten-von Mölken M, Bakker C, van Doorslaer E, van der Linden Sj. Methodological issues of patient utility measurement: experience from two clinical trials. Med Care. 1995;33: 922-37.
17. Dubourg WR, Jones-Lee MW, Loomes G. Imprecise preferences and the WTP-WTA disparity. J Risk Uncertainty. 1994; 9:115-33.
18. Kahnemann D, Tversky A. Prospect theory: an analysis of decision-making under risk. Econometrica. 1979;47:263-91.
19. Altman DG. Practical Statistics for Medical Research. London, U.K.: Chapman and Hall, 1991.
20. Amemiya T. Studies in Econometric Theory. Aldershot, U.K.: Edward Elgar, 1994.

APPENDIX A

Instructions for the Time-tradeoff Questions

In the **time-tradeoff method you are confronted with a choice between two health profiles:**

X years in a specific health state followed by death

Y years in full health followed by death

The value for X has been given. You are requested, given this value of X, to indicate on a line for what value(s) of Y you consider the two profiles to be equivalent. For example, if you consider *20* (=X) years **in health state A to be equivalent to 15 years in full health, then your Y value is equal to 15.**

One way to answer the time-tradeoff questions is by indicating with a – sign those values of Y for which you definitely prefer the first profile (X years in the given health state) and with a + sign those values of Y for which you definitely prefer the second profile (Y years in full health). Finally, indicate with an * sign those values of Y for which you find it hard to choose between the profiles.

APPENDIX B

Instructions for the Standard-Gamble Questions

A standard gamble consists of two alternatives:

X years in a specific health state for certain followed by death

Treatment with two possible outcomes. If treatment is successful you will be in full health for X years followed by death. If treatment fails you will die immediately.

You are requested to indicate on a line with probabilities of successful treatment p for which value of p you consider the two alternatives to be equivalent. For example, if **you** consider treatment with a probability of success of 60% to be equivalent to X years in the specific health state for certain then p is equal to 60%.

One way to answer the standard-gamble question is to indicate with a – sign those values of p for which you definitely prefer the certain health state and with a + sign those values of p for which you definitely prefer the treatment option. Finally indicate with a * sign those values of p for which you find it hard to choose between the two profiles.

Next to the line with probabilities of successful treatment, a line has been drawn that shows the corresponding probabilities of failure of treatment. This has been done in order to remind you what your choices imply in terms of the probability of failure of treatment.