

Standard gamble, time trade-off and rating scale: experimental results on the ranking properties of QALYs

Han Bleichrodt ^{a,*}, Magnus Johannesson ^b

^a *Institute for Medical Technology Assessment (iMTA), Erasmus University, PO Box 1738, 3000 DR, Rotterdam, The Netherlands*

^b *Centre for Health Economics, Stockholm School of Economics, Stockholm, Sweden*

Received 1 April 1995; accepted 1 May 1996

Abstract

This paper compares the relative performance of quality adjusted life years (QALYs) based on quality weights elicited by rating scale (RS), time trade-off (TTO) and standard gamble (SG). The standard against which relative performance is assessed is individual preference elicited by direct ranking. The correlation between predicted and direct ranking is significantly higher for TTO-QALYs than for RS-QALYs and SG-QALYs. This holds both based on mean Spearman rank correlation coefficients calculated per individual and based on two social choice rules: the method of majority voting and the Borda rule. Undiscounted TTO-QALYs are more consistent with direct ranking than discounted TTO-QALYs.

JEL classification: C91; I10

Keywords: Cost utility analysis; Health state utility measure; Individual preferences

1. Introduction

Cost-effectiveness analysis in which costs are measured in monetary units and health effects are measured in non-monetary units is at the moment the most

* Corresponding author.

common approach to carry out economic evaluations of health care programmes. To measure the effects of different medical interventions in a comprehensive way, an outcome measure is needed that simultaneously takes into account quality of life and quantity of life. Quality adjusted life years (QALYs) have been proposed as a measure that can accommodate this requirement. Cost-effectiveness analysis in which QALYs are used as the outcome measure is generally referred to as cost utility analysis (Drummond et al., 1987).

QALYs are calculated by adjusting life years for the quality of life in which they are spent. Health states are assigned a quality weight that lies between 0 and 1. Three principal methods exist to estimate these quality weights based on the preferences of individuals: ¹ the rating scale (RS), the time trade-off (TTO) and the standard gamble (SG). ² Empirical studies have produced evidence that these three methods elicit different quality weights, the general pattern being that, given common scaling, the standard gamble elicits higher quality weights than the time trade-off which in turn elicits higher weights than the rating scale (see, for example, Torrance (1976), Wolfson et al. (1982), Read et al. (1984) and Hornberger et al. (1992)). The worrying implication of these findings is that QALY based decision making may lead to different policy recommendations depending on which of the three methods is used to elicit the quality weights.

No consensus currently exists as to which of the three methods should be preferred. Several authors have argued that from a theoretical point of view the standard gamble is the preferred method (e.g. Weinstein and Fineberg, 1980; Torrance and Feeny, 1989). The standard gamble has a well-established axiomatic foundation, being an appropriate method to measure von Neumann Morgenstern expected utilities. This point of view has been disputed by Broome (1993) and Richardson (1994), among others. Richardson argues firstly that the axioms underlying expected utility theory are empirically flawed, and secondly that the theoretical basis for expected utility theory is defective. In Richardson's opinion, the time trade-off comes closest to four criteria that are necessary to ensure that a measurement unit satisfies the purported objective of QALYs: the combination of quantity of life and quality of life into a single measure that can be used in cost utility analysis. Broome on the other hand argues that both standard gamble and time trade-off are unnecessarily restrictive in terms of the individual preference relation, and that for this reason the rating scale should be the preferred method.

Empirical studies comparing the three assessment methods can be broadly divided into two categories: studies that take one of the methods (typically the standard gamble) as the standard against which the performance of the other

¹ In this paper we use the term "quality weights." Other terms are in use as well, for example "preference scores." In the context of this paper these terms are equivalent.

² There exist other approaches to estimate quality weights for health states, for example Nord (1995) has used and advocated the person trade-off technique.

methods is judged and studies that only compare the quality weights elicited by the three methods without drawing any inferences about their relative performance.

This paper is rooted in decision theory. The purpose of decision theory is to explain individual preference relations (Fishburn, 1970; Wakker, 1989). Axiomatizations are aimed at making the individual preference relations tractable by means of a model. In this paper we take individual preference relations with respect to health profiles as the basic data to be explained. QALYs are considered to be a model to explain individual preferences concerning health profiles.³ In order to interpret QALYs as a model several restrictive assumptions have to be imposed (see, for example, Pliskin et al. (1980); Broome (1993); Bleichrodt (1995)). Depending on which method is used to measure the quality weights, three types of models can be distinguished: (i) QALYs based on standard gamble weights (SG-QALYs); (ii) QALYs based on time trade-off weights (TTO-QALYs); and (iii) QALYs based on rating-scale weights (RS-QALYs). The aim of the present study is to examine by means of an experiment which of these three models corresponds most closely to individual preferences, measured by the direct ranking of health profiles. The structure of the paper is as follows. In Section 2 we briefly describe the standard gamble, the time trade-off and the rating scale. Section 3 describes the design of the experiment by means of which we aim to test the correspondance of SG-QALYs, TTO-QALYs and RS-QALYs with the direct ranking of health profiles. Section 4 contains a description of the analytical methods used in the paper. The results are presented in Section 5. Section 6 contains concluding remarks.

2. Standard gamble, time trade-off and rating scale

Extensive discussions of the standard gamble, the time trade-off and the rating scale can be found elsewhere in the literature (Torrance, 1986; Drummond et al., 1987). Here we confine ourselves to a concise description.

In the standard gamble method quality weights for health states are determined by comparing a specific number of years in health state Q for certain with a gamble (treatment) offering two reference outcomes, which are typically full health for the same number of years and immediate death. The probability (p) of

³ It is important to note that this interpretation of QALYs as a utility model based on individual preferences, which underlies this paper, is not shared by all authors in the field. According to one line of research represented by, for instance, Williams (1985), the trade-off between quality and quantity of life is a socio-political question and QALYs need not necessarily reflect individual preferences. Nord (1994) has also argued in favour of using QALYs as a measure of social value rather than individual utility.

full health is varied until the respondent is indifferent between the two alternatives. This indifference probability is the weight to be assigned to health state Q .

The time trade-off method, developed by Torrance et al. (1972), requires a respondent to compare Y years in a particular health state Q to X years in full health. The number X is varied until the respondent is indifferent between the alternatives. The quality weight assigned to health state Q is then set equal to X/Y .

Finally, in the rating scale method a respondent locates the health state(s) to be assessed on a line calibrated from 0 (immediate death) to 100 (full health). The scale is subsequently normalized to immediate death = 0 and full health = 1 and the resulting health state weight is calculated by dividing the rating scale weight by 100.

3. Design of the experiment

The aim of the experiment is to examine the relative performance of SG-QALYs, TTO-QALYs and RS-QALYs in terms of their ability to predict individual preferences over health profiles. Respondents were 80 students from the Stockholm School of Economics and 92 students from the Erasmus University Rotterdam. The students were paid approximately \$15 for their participation in the study. The experiment was carried out in different sessions lasting approximately 1 h with on average 10 individuals per session. The procedure followed in each session was first to explain the task to respondents, then to ask respondents to perform the specific task and then to explain 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.

We selected eight health states to be included in the questionnaire. The health states were taken from the Maastricht Utility Measurement Questionnaire, a slightly adapted version of the McMaster Health Utility Index (Bakker et al., 1994; Rutten-van Mölken et al., 1995). The selected health states correspond to commonly occurring types of back pain and rheumatism. Health states in the Maastricht Utility Measurement Questionnaire consist of six attributes. We excluded two attributes, side effects of medicines and anxiety about prognosis, from the health state descriptions. These attributes were excluded because they were not essential for the purposes of this experiment and because we feared that giving too much information to respondents might unnecessarily complicate the tasks they were faced with. The remaining four attributes were general daily activities, self care, leisure activities and pain. The attributes and the levels of the attributes are shown in Table 1. The health states were indicated by capital letters and were described on a set of cards, which were handed out to respondents at the beginning

Table 1

The multi-attribute health status classification system used in the experiment

General daily activities

- (1) Able to perform all tasks at home and/or at work without problems
- (2) Able to perform all tasks at home and/or at work, albeit with some difficulties
- (3) Not able to perform some tasks at home and/or at work
- (4) Not able to perform many tasks at home and/or at work
- (5) Not able to perform any task at home and/or at work

Self care

- (1) Able to perform all self care activities (eating, washing, dressing) without problems
- (2) Able to perform all self care activities (eating, washing, dressing), albeit with some difficulties
- (3) Not able to perform some self care activities (eating, washing, dressing)
- (4) Not able to perform many self care activities (eating, washing, dressing) without help
- (5) Not able to perform any self care activity (eating, washing, dressing) without help

Leisure activities

- (1) Able to perform all types of leisure activities without difficulties
- (2) Able to perform all types of leisure activities, albeit with some difficulties
- (3) Not able to perform certain types of leisure activities
- (4) Not able to perform many types of leisure activities
- (5) Not able to perform any type of leisure activities

Pain and/or other complaints

- (1) No pain and/or other complaints
 - (2) Now and then light to moderate pain and/or other complaints
 - (3) Often light to moderate pain and/or other complaints
 - (4) Often moderate to severe pain and/or other complaints
 - (5) Always severe pain and/or other complaints
-

of each session. Health state D , which is relevant for the analysis of this paper is described in Table 2.

We divided the questionnaire into different sections. The first substantive task respondents were confronted with was the ranking of six of the health states in terms of desirability to themselves. For reasons not related to the present study the 6 health states to be ranked varied per session. However, for every respondent health state D was included in the ranking task. After the ranking task respondents were asked to locate the health states on a rating scale. In the third section respondents answered three time trade-off questions. All respondents answered a

Table 2

Description of health state D used in the experiment

-
- 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 activity
 - Often moderate to severe pain and/or other complaints
-

question where one of the alternatives was 30 years in health state *D*. The answer to this time trade-off question is used in the subsequent analysis of this paper as the time trade-off quality weight for health state *D*. Value elicitation was on a line with numbers of years in full health. Respondents were encouraged to indicate first the values of *X*, the number of healthy years, for which they definitely preferred 30 years in health state *D*, then the values of *X* for which they definitely preferred *X* years in full health and finally those values of *X* for which they found it hard to choose between the alternatives. Respondents were explained that they could indicate a range of values for *X* for which they found it hard to choose, but they were encouraged to make this range as small as possible. This format was adopted to allow respondents to express imprecision of preferences (Dubourg et al., 1994). Trading off life years is a task respondents are relatively unfamiliar with and their preferences may be somewhat imprecise. In our format we attempted to take this imprecision of preference into account. For individuals who indicated a range of values for *X*, we used the mid-point of this interval as their time trade-off quality weight for health state *D*.

Section four consisted of three standard gamble questions. All versions of the questionnaire contained a question where 30 years in health state *D* was the certain option. Respondents' answers to this question were used in the analysis. 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: only displaying the probability of successful treatment might cause individuals to focus on successful treatment, not sufficiently taking into account the probability of failure of treatment. Psychological evidence on the influence of reference effects on choice is abundant (e.g. Tversky and Kahneman, 1991). Similar to the time trade-off question, an attempt was made to take imprecision of preferences into account. Respondents were asked first 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. For individuals who indicated a range of values for *p*, we used the mid-point of this interval as their standard gamble quality weight.

In section five the respondents were asked to rank seven health profiles, i.e. combinations of quality of life and quantity of life, in terms of desirability to themselves. The ranking exercise was intended to measure individual preferences for health states directly. The health profiles were printed on a set of cards which were distributed together with the questionnaire and the set of health states. This ranking task was similar to the task in section one, except that the objects to be ranked were health profiles rather than health states. The seven health profiles are described in Table 3. The seven health profiles differed over 20 years. After 20 years all profiles resulted in death. Profiles lasting less than 20 years resulted in

Table 3
The seven health profiles included in the experiment

Number profile	Years in full health	Years in <i>D</i>	Years dead
1	0	20	0
2	18	0	2
3	16	0	4
4	14	0	6
5	12	0	8
6	8	8	4
7	6	11	3

earlier death. In the case of mixed health profiles consisting both of years in full health and years in *D*, the years in full health always came first. It was expected, and confirmed in the pilot sessions, that profiles of decreasing quality of life are more intuitive to respondents than profiles of increasing quality of life.

To be able to compare the relative performance of SG-QALYs, TTO-QALYs and RS-QALYs in terms of the direct ranking of the health profiles, health profiles are needed that will be ranked differently if the QALY weight differs between the methods. The seven health profiles were therefore selected with the intention in mind to produce different rankings for SG-QALYs, TTO-QALYs and RS-QALYs. Table 4 uses hypothetical weights between 0.5 and 1.0 to show how the implied QALY ranking of the profiles differs for different quality weights. The experiment was designed with a time trade-off quality weight of about 0.7 in mind. Because empirical evidence has indicated that the standard gamble generally results in higher quality weights than the time trade-off which in turn gives higher weights than the rating scale, we hoped to create the conditions under which the implied rankings for SG-QALYs, TTO-QALYs and RS-QALYs were likely to differ. In the pilot test the average time trade-off weight for health state *D* was about 0.69.

Table 4
Implied QALY rankings of the seven health profiles for different quality weights with no discounting¹

Number profile	Weight = 0.5 rank	Weight = 0.6 rank	Weight = 0.7 rank	Weight = 0.8 rank	Weight = 0.9 rank	Weight = 1.0 rank
1	7	6	3	2	1	1
2	1	1	1	1	1	2
3	2	2	2	2	3	4
4	3	3	3	6	6	6
5	4	6	7	7	7	7
6	4	4	6	5	5	4
7	6	5	5	4	4	3

¹ When two profiles are given the same ranking (e.g. profiles 1 and 2 with a weight of 0.9) this signifies a tie (i.e. the number of QALYs is the same for both profiles).

The standard gamble weight was 0.76 and the rating scale weight was 0.36. Health state *D* appeared to be a good candidate to test the ranking properties and was therefore selected. Note that the ranking of some of the health profiles in Table 4 will by definition be the same for all the QALY measures. Profiles 2–5 for instance consist of profiles of varying duration in full health followed by death. We included profiles 2–5 in the experiment to ensure variation in the ranking of all profiles over a rather wide range of quality weights. Without enough variation in the rankings over the range of quality weights we might not detect differences between the methods in correlation with the direct ranking. If the three methods resulted in different quality weights, but the rankings did not vary, then over this range of quality weights our analysis would not be informative.

In economic evaluations of health care programmes costs and effects are generally adjusted for their timing, by discounting at a fixed rate (see Gafni and Torrance (1984) for an analysis of time preference in health). To examine to what extent discounted QALY maximization models are consistent with the individual preference relation, rankings were compared for various discount rates. Discounting implies that life years get different weights. Johannesson et al. (1994) have shown that in the case of discounting, in order to be consistent with individual preferences, the time trade-off quality weights have to be adjusted for discounting. This is achieved by discounting the equivalent number of years in full health and the 30 years in health state *D* before the quality weight is derived. Johannesson et al. also argued that the standard gamble weights are not affected by discounting, because the time horizon is the same for the assessed health state and for full health. The intuition behind their argument is as follows. The validity of standard gamble weights for health states depends on a preference condition which has to be imposed on top of the von Neumann and Morgenstern axioms: utility independence of health status from time duration. This means that utilities for health states can be assessed holding the time duration constant both for the certain outcome and for the gamble outcomes. Given a common time duration, even if individuals apply a positive rate of discount to life years, this will affect all outcomes in a similar way and thus the standard gamble weights are not affected by positive discounting.

The rating scale quality weight is elicited without reference to time duration and therefore no adjustment for discounting is necessary.

4. Methods of analysis

The seven health profiles were translated for each respondent to SG-QALYs, TTO-QALYs and RS-QALYs on the basis of the elicited quality weights. The predicted rankings of the health profiles were then compared with the direct ranking which was elicited in section five of the questionnaire. We examined the predictive power of SG-QALYs, TTO-QALYs and RS-QALYs both at the indi-

vidual and at the societal level. At the individual level, we compared for each individual the predicted ranking of the health profiles by each of the three models with the direct ranking. To assess the strength of the association we calculated for each individual and for each method the Spearman rank correlation coefficient between the predicted ranking and the direct ranking. These rank correlation coefficients were then averaged over all individuals. The Spearman rank correlation coefficient is a non-parametric technique which is applicable to ordinal data. Given that the direct ranking data were ordinal, parametric correlation coefficients could not be applied. The Spearman rank correlation coefficient lies between -1 and 1 , a higher value indicating stronger positive association between the ranks, a value of zero indicating no association. The QALY measure with the highest average Spearman rank correlation coefficient is most closely associated with the direct ranking of the health profiles.

For the analysis at the societal level we had to aggregate individual preferences into social preference. A problem arises here because of the nature of our data. The direct ranking only provides information with respect to the ordering of profiles. No cardinal information is available. Arrow (1951, 1963) has proved that it is impossible to construct a social ordering from individual orderings, that satisfies four “very mild looking conditions” (Sen, 1970)⁴. We examined two simple social choice rules, each violating one of Arrow’s conditions. First, we applied the method of majority voting. Ranking one profile above another was interpreted to be a vote in favour of the former. We constructed a social preference relation from an examination of the votes between every possible pair of profiles. A problem with the social preference relation thus constructed is that it need not be transitive. Therefore we also constructed a social preference relation based on the Borda rule. The Borda rule assigns points to profiles corresponding to the ranks of the profiles and then sums these points over all individuals. The points assigned to a profile were set equal to the rank of the profile. That is, we assigned points in descending order, i.e. a lower number meaning “more preferred.” The Borda rule satisfies transitivity, but violates the condition Arrow refers to as “independence of irrelevant alternatives.” According to this condition social preference between two alternatives should not be affected by a third alternative. The method of majority voting and the Borda rule each satisfy a different subset of Arrow’s conditions. The union of these sets consists of Arrow’s conditions.

In interpreting the results on majority voting it is important to realize that the two exercises of ranking health profiles in terms of their desirability for the individual and of voting between health profiles, which we set equal, may in fact not be equivalent. In a voting situation the individual may consider the desirability of the alternatives both for themselves and for others whereas in the ranking task

⁴ The four conditions most frequently referred to are unrestricted domain, weak Pareto principle, independence of irrelevant alternatives and non-dictatorship.

respondents were asked to consider the alternatives in terms of desirability to themselves (see Labelle and Hurley (1992) for a discussion of the potential importance of preferences over outcomes for others).

It is worth emphasizing here that the analysis of the results we present in Section 5 is based on statistical significance. Statistical significance need not necessarily equal economic significance. That is, observed statistically significant differences may not be meaningful. We discuss the difference between statistical significance and meaningfulness in more detail in Section 6.

5. Results

5.1. Disparity between the methods

Differences in ranking performance can only occur if the three methods produce different weights. Table 5 shows that the mean quality weights for standard gamble, time trade-off and rating scale indeed differ significantly. Differences are significant at the 0.1% level, both for the Dutch, the Swedish and the total sample. The difference between the Dutch and the Swedish sample is not significant ($p > 0.10$) for the time trade-off and the standard gamble quality weights. However, for the rating scale the difference is significant at the 10% level ($p \approx 0.09$). Compared to the results of the pilot study the standard gamble weights and the time trade-off weights are approximately 0.10 lower, whereas the rating scale weight is slightly higher.

5.2. Spearman rank correlation coefficients

Table 6 displays the mean of the Spearman rank correlation coefficients between the direct rankings and the rankings predicted by RS-QALYs, TTO-QALYs and SG-QALYs, respectively. Both in the Dutch and in the Swedish sample TTO-QALYs are significantly stronger correlated with the direct ranking than RS-QALYs and SG-QALYs. The difference between RS-QALYs and SG-

Table 5

Mean SG, TTO and RS quality weights (preference scores) for health state *D*. Standard errors within parentheses

Method	Dutch sample	Swedish sample	Total sample	Difference, NL – S
RS	0.3867 ¹ (0.0149)	0.4274 ¹ (0.0191)	0.4056 ¹ (0.0199)	$p = 0.092$
TTO	0.5958 ^{1,2} (0.0238)	0.5575 ^{1,2} (0.0237)	0.5780 ^{1,2} (0.01684)	n.s.
SG	0.6786 ² (0.0279)	0.6620 ² (0.0257)	0.6709 ² (0.0191)	n.s.

¹ Significantly different from SG at 99% confidence level.

² Significantly different from RS at 99% confidence level.

Table 6

Mean Spearman rank correlation coefficients between the direct ranking and the predicted rankings of RS-QALYs, TTO-QALYs and SG-QALYs. Standard errors within parentheses. No discounting

Comparison	Dutch sample	Swedish sample	Total sample	Difference, NL – S
RS—direct	0.7208 (0.033)	0.7932 (0.027)	0.7545 (0.022)	$p = 0.088$
TTO—direct	0.8194 ^{1,2} (0.027)	0.8669 ^{1,2} (0.018)	0.8415 ^{2,3} (0.017)	n.s.
SG—direct	0.6891 (0.036)	0.7684 (0.027)	0.7259 (0.023)	$p = 0.062$

¹ Significantly different from RS at 95% confidence level.

² Significantly different from SG at 99% confidence level.

³ Significantly different from RS at 99% confidence level.

QALYs is not significant, even though the mean rank correlation coefficient is higher for RS-QALYs in both samples.

The Swedish responses are more consistent with the given direct ranking than the Dutch responses for each of the three methods. The difference in mean rank correlation coefficient between the Swedish and the Dutch responses is significant at the 90% confidence level for the rating scale and for the standard gamble, but it is not significant for the time trade-off.

Table 6 was constructed under the assumption of no discounting. Table 7 shows the results for a discount rate of 5%. This is the situation most frequently encountered in cost utility analyses.

A comparison between Tables 6 and 7 reveals that 5% discounting reduces the rank correlation coefficients for rating scale and time trade-off and increases the rank correlation coefficient for the standard gamble. Even though the difference is less significant, the rank correlation with the direct ranking is still higher for TTO-QALYs than for RS-QALYs and SG-QALYs. For the latter two, the difference between the mean rank correlation coefficients is not significant.

A comparison between Tables 6 and 7 shows that for the standard gamble the mean rank correlation coefficient increases with 5% discounting. The difference between the mean rank correlation coefficients is highly significant ($p < 0.001$).

Table 7

Mean Spearman rank correlation coefficients between the direct ranking and the predicted rankings of RS-QALYs, TTO-QALYs and SG-QALYs. Standard errors within parentheses. Five per cent discounting

Comparison	Dutch sample	Swedish sample	Total sample	Difference, NL – S
RS—direct	0.7058 (0.035)	0.7795 (0.030)	0.7401 (0.023)	n.s.
TTO—direct	0.7886 ^{1,2} (0.028)	0.8481 ¹ (0.020)	0.8162 ^{2,3} (0.018)	$p = 0.086$
SG—direct	0.7297 (0.033)	0.8274 (0.026)	0.7752 (0.023)	$p = 0.022$

¹ Significantly different from RS at 90% confidence level.

² Significantly different from SG at 90% confidence level.

³ Significantly different from RS at 95% confidence level.

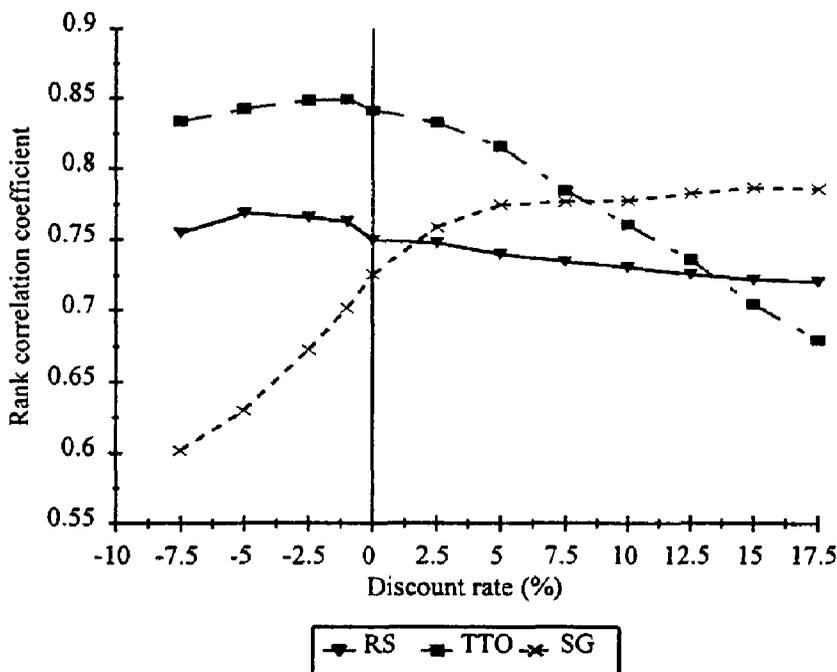


Fig. 1. The relationships between discount rates and rank correlation coefficients. The figure displays the relationship for RS-QALYs, TTO-QALYs and SG-QALYs between the mean Spearman correlation coefficient and the discount rate used. Positive discounting makes QALY-based decision making more consistent with individual preferences elicited by direct ranking for the standard gamble, but less consistent for the time trade-off and the rating scale.

Fig. 1 shows for the total sample that for discount rates higher than 5%, the rank correlation coefficient for SG-QALYs increases even more. For the total sample the maximizing discount rate for SG-QALYs is 15.5% ($rcc = 0.789$). For the Swedish sample the maximum rank correlation coefficient for SG-QALYs is attained at a discount rate of approximately 9% ($rcc = 0.835$). For the Dutch sample the maximum rank correlation coefficient for SG-QALYs is attained at a discount rate of approximately 16% ($rcc = 0.760$).

For the rating scale and the time trade-off we observe the opposite pattern: 5% discounting has a decreasing impact on the mean rank correlation coefficient. For the rating scale the difference is significant at the 5% level. For the time trade-off the difference is significant at the 0.1% level. Fig. 1 shows that the maximizing discount rate for RS-QALYs is approximately -5% ($rcc = 0.769$) for the total sample. For the Swedish and the Dutch samples the maximizing discount rates for RS-QALYs are -1% ($rcc = 0.804$) and -5% ($rcc = 0.756$), respectively. For TTO-QALYs the maximizing discount rate is approximately -1% ($rcc = 0.850$). For the Dutch and Swedish samples the maximizing discount rates are -5% ($rcc = 0.829$) and -1% ($rcc = 0.876$), respectively.

5.3. Majority voting

Table 8 summarizes the evidence on majority voting. Profiles are shown in decreasing order of preference according to majority voting. The first column shows the results for the direct ranking exercise. Columns two to four show the results for the three methods with no discounting and columns five to seven show the results for the three methods with 5% discounting. The number in parentheses shows the percentage of respondents who favour the profile over the profile coming next in preference. For example, the column “Direct” shows that in the direct ranking task profile 2 came out as most preferred, profile 3 as second most preferred, profile 4 as third most preferred, etc. Profile 2 was preferred by all respondents to profile 3, profile 3 was preferred by all respondents to profile 4, and profile 4 was preferred by 71% of respondents to profile 6.

Table 8 shows that for no discounting, the ranking predicted by TTO-QALYs corresponds most closely to the ranking being given: the predicted rank order of the profiles is similar to the rank order elicited in the direct ranking task. Moreover, the predicted percentage of respondents “voting” in favour of a profile is in most cases quite similar to the percentage of respondents “voting” in favour of a profile according to direct ranking. To obtain insight in the correspondence of the proportions voting in favour of a particular profile, we calculated the correlation coefficients between the proportion of votes in favour of profile i over profile j based on the direct ranking and the predicted proportion of votes in favour of

Table 8

Majority vote ordering of profiles. The direct ranking of profiles is shown in the first column and the following columns show the predicted ranking of profiles based on RS-QALYs, TTO-QALYs and SG-QALYs with 0 and 5% discounting. The proportions of individuals preferring a profile to the next profile in the ordering are shown within parentheses¹

Direct	RS (0%)	TTO (0%)	SG (0%)	RS (5%)	TTO (5%)	SG (5%)
2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)
3 (100%)	3 (100%)	3 (100%)	3 (64%)	3 (100%)	3 (100%)	3 (100%)
4 (71%)	4 (100%)	4 (75%)	1 (54% vs. 59%)	4 (100%)	4 (55%)	4 (60%)
6 (84%)	5 (75%)	6 (62%)	4 = 7 (55% vs. 59%)	5 (78%)	6 (55%)	6 (59%)
7 (52%)	6 (95%)	7 (56%)		6 (98%)	7 (58%)	7 (66%)
5 (53%)	7 (95%)	5 (50.3%)	6 (79%)	7 (99%)	1 (56%)	1 (51%)
1	1	1	5	1	5	5
ρ	0.902	0.988	0.859	0.868	0.951	0.970

¹ For SG-QALYs with no discounting profiles 4 and 7 were given the same ordering according to the majority voting rule (i.e. the proportion of respondents “voting for” 4 over 7 was exactly 50%). The 54% vs. 59% given in parentheses means that 54% were predicted to prefer profile 1 to profile 4 and 59% were predicted to prefer profile 1 to profile 7. The 55% vs. 59% given in parentheses similarly means that 55% were predicted to prefer profile 4 to profile 6 and 59% were predicted to prefer profile 7 to profile 6.

profile i over profile j based on each of the three methods. These correlation coefficients are shown in the bottom rows of Table 8.⁵ In the situation of no discounting the correlation coefficient is significantly higher for the time trade-off than for the rating scale and the standard gamble at the 0.1% significance level.⁶ The difference between rating scale and standard gamble is not significant. In the situation of 5% discounting the correlation coefficient is highest for the standard gamble. Only the difference between standard gamble and rating scale is significant ($p = 0.022$).

Table 8 also suggests an explanation why SG-QALYs and RS-QALYs are less consistent with the results of the direct ranking exercise than TTO-QALYs. The explanation we suggest is that the standard gamble assigns too much weight to health state D compared to the weight implied by direct ranking. The rating scale does not assign enough weight to health state D compared to the weight implied by direct ranking. To illustrate the first claim made, that the standard gamble assigns too much weight compared to direct ranking, compare for example profiles 1 and 5. The difference between profiles 1 and 5 is that profile 1 offers more life years than profile 5, but these life years are spent in a lower quality of life. Profile 1 will be preferred to profile 5 if the utility gain of 8 additional years in health state D more than compensates the utility loss of spending the first 12 years in health state D rather than in full health. For example, in the situation of no discounting, profile 1 will be preferred to profile 5 if $8 * [U(D) - U(\text{immediate death})] > 12 * [U(\text{full health}) - U(D)]$.⁷ Or, given the scaling of U , if $U(D) > 0.6$. Table 8 shows that according to the standard gamble profile 1 is preferred to profile 5 by a majority of respondents. In fact 70.9% preferred profile 1 to profile 5 according to SG-QALYs. Thus 70.9% of the respondents assigned a quality weight greater than 0.6 to health state D by the standard gamble. However, in the direct ranking a majority of respondents (53%) ranked profile 5 above profile 1, which implies that only 47% of the respondents assigned a quality weight greater than 0.6 to health state D in the direct ranking task. Thus, compared to the direct ranking the standard gamble gives too much weight to health state D .

Positive discounting has the effect of reducing the utility gain of profile 1 over profile 5 relative to the utility loss. For example with 5% discounting the utility gain is $3.78 * U(D)$ and the utility loss is $9.31 * [1 - U(D)]$. Profile 1 will now be preferred to profile 5 if $U(D) > 0.71$. Those respondents who assigned a standard gamble weight between 0.6 and 0.71 will now prefer profile 5 to profile 1. Thus,

⁵ Obviously if the pair (i, j) , i.e. the proportion of voters favouring profile i over profile j , was included in the calculation (j, i) was not. The pair (j, i) is the complement of (i, j) and therefore no new information is added to the analysis by including (j, i) .

⁶ In testing for significance use was made of Fisher's Z-transformation: $Z_F = 0.5 * \ln[(1 + \rho)/(1 - \rho)]$.

⁷ Note that this analysis assumes intertemporal separability of individual preferences. However, this is an assumption that has to be made to characterize QALYs as a utility model.

Table 9

Correlation coefficients for the two samples for the majority vote ordering of profiles

Sample (% discounted)	RS	TTO	SG
NL (0)	0.879	0.979 ^{1,2}	0.806
S (0)	0.924	0.982 ^{2,3}	0.862
NL (5)	0.856	0.932	0.928
S (5)	0.894	0.948	0.981 ¹

¹ Significantly different from RS at 99% confidence level.² Significantly different from SG at 99% confidence level.³ Significantly different from RS at 95% confidence level.

positive discounting has the effect of making SG-QALYs more consistent with direct ranking by decreasing the number of respondents who are predicted to prefer profile 1 to profile 5.

The hypothesis that the rating scale assigns too low a weight to health state D compared to the weight implied by the direct ranking can be seen for example by comparing profiles 5 and 6. With no discounting a similar calculation exercise as above shows that profile 6 will be preferred to profile 5 if $U(D) > 0.5$. The direct ranking exercise revealed that a majority of respondents preferred profile 6 to profile 5 (64.6%). That is, 64.6% of the respondents assigned a weight greater than 0.5 to health state D in the direct ranking task. However, in the rating scale task only 25% of the respondents assigned a weight greater than 0.5 to health state D . Obviously, for the rating scale positive discounting only makes things worse. With 5% discounting profile 6 will be preferred to profile 5 if $U(D) > 0.55$. Thus respondents who give a quality weight between 0.50 and 0.55 to health state D are now predicted by RS-QALYs to prefer profile 5.

Table 9 displays the results for the Swedish and Dutch samples separately. We observe a similar pattern as for the total sample. Correlation coefficients are without exception higher in the Swedish sample. However, the difference between the two samples is only significant for SG(5%) ($p = 0.000$).

5.4. Borda rule

Table 10 shows social preferences according to the Borda rule for the total sample. Recall from Section 4 that in a comparison of profiles, the profile with the lowest score is preferred. Thus for the direct ranking exercise the preference order implied by the Borda rule is $2 > 3 > 4 > 6 > 7 > 1 > 5$.

Table 10 shows that in the situation where no discounting is applied the ranking predicted by TTO-QALYs is most closely related to the direct ranking. Table 10 confirms the pattern we already observed with respect to Table 8. The standard gamble assigns too high a quality weight to health state D compared to the weight

Table 10

Mean Borda scores for the health profiles. The first column shows the Borda scores based on the direct ranking of profiles and the following columns shows the Borda scores based on the predicted ranking of profiles for RS-QALYs, TTO-QALYs and SG-QALYs with no discounting and 5% discounting

Profile	Direct	RS (0%)	TTO (0%)	SG (0%)	RS (5%)	TTO (5%)	SG (5%)
1	5.31	6.69	4.99	3.88	6.91	4.84	4.97
2	1.03	1	1.05	1.16	1	1.06	1.11
3	2.17	2.01	2.21	2.49	2.01	2.36	2.40
4	3.82	3.08	3.81	4.43	3.05	4.33	4.20
5	5.55	4.55	5.63	6.22	4.33	6.06	5.94
6	4.42	4.77	4.77	4.88	4.80	4.62	4.63
7	5.24	5.75	5.04	4.58	5.90	4.71	4.75
ρ_{total}		0.923	0.992	0.895	0.901	0.970	0.980
ρ_{NL}		0.905	0.990	0.874	0.885	0.962	0.959
ρ_{S}		0.941	0.993	0.916	0.917	0.977	0.992

implied by the direct ranking exercise. Positive discounting mitigates this effect: with 5% discounting SG-QALYs more closely reflect the direct ranking. The rating scale assigns too low a weight to health state *D* compared to the weight implied by the direct ranking exercise. Positive discounting reinforces this effect.

The final three rows of Table 10 report for the total sample, the Dutch sample and the Swedish sample the correlation coefficients of the Borda scores assigned by the three methods with the Borda scores calculated on the basis of the direct ranking.⁸ These confirm the remarks made above: the scores predicted by TTO-QALYs are most closely related to the scores resulting from the direct ranking in the case of no discounting. In the case of 5% discounting, the scores predicted by SG-QALYs are most closely related to the scores predicted by direct ranking.

6. Concluding remarks

The aim of this study was to compare RS-QALYs, TTO-QALYs and SG-QALYs in terms of their ability to predict individual preferences over health profiles. The reason we compared the predictions of the three models with individual preferences is that the latter are the basic data that decision theory seeks to explain. Individual preferences were measured by direct ranking of a number of health profiles.

⁸ No significance is reported here. For less than 10 observations the test based on Fisher's Z-transformation is not sufficiently accurate.

The results of the experiment reported in this paper show that, in the situation of no discounting, the correlation between predicted ranking and direct ranking was significantly higher for TTO-QALYs than for RS-QALYs and SG-QALYs. This result held both in terms of average Spearman rank correlation coefficients calculated for each individual and in terms of two social choice rules each satisfying a different subset of Arrow's "reasonable conditions." No significant differences were observed between RS-QALYs and SG-QALYs, though in general RS-QALYs were slightly more consistent with the direct ranking of profiles.

The most common procedure in economic evaluations is to discount QALYs at a fixed rate, generally 5%. With a 5% discount rate the correlation between the predicted ranking and the direct ranking increased for SG-QALYs, but decreased for TTO-QALYs and for RS-QALYs. As we outlined in the previous sections, the reason SG-QALYs more closely reflect the direct ranking may be that the standard gamble assigns a relatively high weight to health state *D* compared to the weight implied by the direct ranking exercise. In the context of our experiment, positive discounting will mitigate this relatively high weight. The suggestion that the standard gamble as it is typically asked in health state valuation, by probability equivalence, results in a relatively high quality weight is consistent with previous findings in the literature (e.g. Hershey and Schoemaker (1985)).

The conclusions described above are entirely based on an analysis of statistically significant differences. It could be argued that even though the correlation between predicted and direct ranking was statistically significantly higher for TTO-QALYs than for RS-QALYs and SG-QALYs, the differences between the methods are not economically important (for a discussion of the distinction between statistical significance and economic importance see McCloskey (1983)). This argument is based on the observation that all methods performed well, because the correlation coefficients were rather high for all three methods by the standards usually used to judge the size of correlation coefficients (cf. Landis and Koch, 1977).

However, one can object to the above argument. It is difficult to interpret the absolute size of the correlation coefficients in the context of this study and to judge the size of the correlation coefficients by the usual standards. The reason is that the predicted ranking of some of the profiles will by definition be the same as the direct ranking as long as individual preferences satisfy monotonicity with respect to years in full health (i.e. preferences between profiles 2, 3, 4 and 5 are obvious) and as long as the individual prefers years in full health to years in health state *D* (i.e. profile 2 will always be preferred to profiles 6 and 7 and profile 3 will always be preferred to profile 6). To illustrate the impact on the size of the correlation coefficients of profiles for which the ranking is obvious, we redid the analysis using only the two combinations of health profiles for which the ordering was not obvious beforehand. These combinations are profiles 1, 4, 6 and 7 and profiles 1, 5, 6 and 7. For the analysis including only profiles 1, 4, 6 and 7 the mean Spearman rank correlation coefficient between predicted and direct ranking

is 0.55 for TTO-QALYs, 0.46 for RS-QALYs and 0.28 for SG-QALYs.⁹ Translated to the classification scheme of Landis and Koch (1977) these rank correlation coefficients would classify as “moderate” for TTO-QALYs and for RS-QALYs and as “fair” for SG-QALYs. For the analysis including only the profiles 1, 5, 6 and 7 the mean Spearman rank correlation coefficient between predicted and direct ranking is 0.54 for TTO-QALYs, 0.30 for RS-QALYs and 0.35 for SG-QALYs.¹⁰ According to the classification scheme of Landis and Koch (1977) the correlation is “moderate” for TTO-QALYs and “fair” for RS-QALYs and for SG-QALYs. The pattern is for both combinations quite similar to the pattern we observed when all health profiles were included in the analysis except that the mean rank correlation coefficients are lower. The differences between the mean rank correlation coefficients are larger and the correlation coefficients do not all fall in the same class according to the classification scheme of Landis and Koch (1977). If one is prepared to adopt the classification scheme suggested by Landis and Koch, which is based on the idea that differences of about 0.20 are meaningful, then the results of this paper suggest that the differences between the methods may not only be statistically significant, but also meaningful. It is, however, difficult to determine exactly what difference in correlation coefficients is economically meaningful. That is, how close is close? This may be an issue for future debate. It should be emphasized once again that this paper has only shown that the correlation between the implied rankings and the direct rankings is statistically significantly higher for TTO-QALYs than for RS-QALYs and for SG-QALYs.

As far as we know, this study is the first which compares the performance of RS-QALYs, TTO-QALYs and SG-QALYs in terms of direct ranking. Apart from offering some tentative conclusions, this paper also raises various questions which may be addressed in future research. First, we measured individual preferences by direct ranking of the seven profiles simultaneously. It is possible that the directly measured rankings have provided a less than ideal assessment of respondents' preferences because of the cognitive burden of the ranking exercise (cf. Fisher, 1979). A different procedure would be to confront individuals with all possible pairs of profiles and to construct a preference ordering from these answers. It is not clear a priori whether the two approaches give identical results. For example, our approach excluded intransitivities. The pairwise approach on the other hand might lead to intransitivities. Second, the approach we used in the time trade-off

⁹ The difference between the mean rank correlation coefficients for TTO-QALYs and for RS-QALYs is not statistically significant. The difference between the mean rank correlation coefficients for TTO-QALYs and for SG-QALYs is significant at the 0.1% level. The difference between the mean rank correlation coefficients for RS-QALYs and for SG-QALYs is significant at the 5% level.

¹⁰ The difference between the mean rank correlation coefficients for TTO-QALYs and for RS-QALYs and between the mean rank correlation coefficients for TTO-QALYs and for SG-QALYs are significant at the 0.1% level. The difference between the mean rank correlation coefficients for RS-QALYs and for SG-QALYs is not statistically significant.

and standard gamble questions is close to the ping-pong approach favoured by many researchers in the field, but it is not exactly similar. Moreover, to accommodate imprecision of preferences we allowed respondents to indicate ranges of values. Although we do not believe that our slightly different procedures have affected the results, it may be worth investigating the sensitivity of the results to this difference in approach. Third, we used only one health state and only a limited number of profiles for which the ranking was not obvious. It may be that the time trade-off is a useful heuristic for a number of health states, but that it does not work equally well for all health states (cf. Stalmeier et al. (1996)). It is worth redoing the analysis using different health states and profiles. Fourth, the fact that we used group sessions rather than individual sessions may have decreased the care with which some individuals answered the questionnaire. This may, in particular, have affected the standard gamble responses. The standard gamble is generally considered to be the most complicated method of the three. On the other hand, as can be seen from Table 5, the pattern of differences in quality weights between the methods is similar to that observed in other studies. Fifth, it is possible that the performance of the methods is affected by the ordering of the tasks in the experiment. All respondents were first asked to perform the rating scale task, then the time trade-off task and finally the standard gamble task. The reason we opted for this order was that in general the rating scale is considered the easiest method to answer and the standard gamble the most complicated. However, it may be that during the experiment respondents became more aware of their “true” preferences and thus the higher consistency of TTO-QALYs over RS-QALYs may simply be a consequence of the order in which the tasks were performed. Even though at the end of the experiment we urged respondents to carefully read through their responses again and to make changes where they thought appropriate, ordering effects may have affected the results. Future experimental studies may wish to randomize the order of the tasks or, alternatively, respondents may be asked to perform only one task.

Finally, two notes of warning are worth making. First, we interpreted QALYs as a utility model. Even though this appears to be the most common interpretation of QALYs (for example the recent debate on the merits of QALYs vs. healthy-years equivalents (Buckingham (1993); Culyer and Wagstaff (1993); Culyer and Wagstaff (1995); Gafni et al. (1993); Loomes (1995); Johannesson (1995); Bleichrodt (1995)) focused on the question of the consistency of QALYs with individual preferences), there are other interpretations, as we remarked before. Second, our results only bear relevance for the descriptive validity of the various QALY models. It may be that for normative/prescriptive reasons, which are more relevant in health economics and medical decision making, one wishes to stick to SG-QALYs. Moreover, if QALYs are intended as decision aids to prescribe individual choices, the paradoxical result emerges that once a model corresponds perfectly with direct choices the model loses its significance for prescriptive purposes. In the case of perfect correspondence one could simply let individuals

choose intuitively and no decision-aiding analysis would contribute anymore. Predictive models can only be of use if they deviate somewhat from actual choice. The question then obviously is how much we allow our measures to deviate from actual choice. This is a question that may be picked up in future research.

Acknowledgements

We would like to thank Eddy van Doorslaer, Peter Wakker and the two anonymous referees for their helpful comments on previous drafts, Maureen Rutten-van Mólken for suggestions with respect to the selection of the health states and Jaco van Rijn for assistance in running the experimental sessions. The experiment reported in this paper was financially supported by the National Corporation of Swedish Pharmacies and by Merck, Sharp and Dhome.

References

- K. Arrow, *Social Choice and Individual Values*, Wiley, New York, 1951, 2nd edn, 1963.
- C. Bakker, M. Rutten-van Mólken, E. van Doorslaer, K. Bennett and S. van der Linden, Feasibility of utility assessment by rating scale and standard gamble in ankylosing spondylitis or fibromyalgia, *J. Rheum.*, 21 (1994) 269–274.
- H. Bleichrodt, QALYs and HYE: under what conditions are they equivalent? *J. Health Econ.*, 14 (1995) 17–37.
- J. Broome, Qalys, *J. Public Econ.*, 50 (1993) 149–167.
- K. Buckingham, A note on HYE, *J. Health Econ.*, 12 (1993), 301–309.
- A.J. Culyer and A. Wagstaff, QALYs versus HYE, *J. Health Econ.*, 12 (1993) 311–323.
- A.J. Culyer and A. Wagstaff, QALYs versus HYE: a reply to Gafni, Birch and Mehrez, *J. Health Econ.*, 14 (1995) 39–45.
- M.F. Drummond, G.L. Stoddart and G.W. Torrance, *Methods for the economic evaluation of health care programmes*, Oxford University Press, Oxford, 1987.
- W.R. Dubourg, M.W. Jones-Lee and G. Loomes, Imprecise preferences and the WTP–WTA disparity, *J. Risk Uncertainty*, 9 (1994) 115–133.
- G.W. Fisher, Utility models for multiple objective decisions: do they accurately represent human preferences? *Decision Sci.*, 10 (1979) 451–479.
- P.C. Fishburn, *Utility Theory for Decision Making*, Wiley, New York, 1970.
- A. Gafni, S. Birch and A. Mehrez, Economics, health and health economics: HYE versus QALYs, *J. Health Econ.*, 12 (1993) 325–339.
- A. Gafni and G.W. Torrance, Risk attitude and time preference in health, *Management Sci.*, 30 (1984) 440–451.
- J.C. Hershey and P.J.H. Schoemaker, Probability versus certainty equivalence methods in utility measurement: are they equivalent? *Management Sci.*, 31 (1985) 1213–1231.
- J.C. Hornberger, D.A. Redelmeier and J. Peterson, Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis, *J. Clin. Epidemiol.*, 45 (1992) 505–512.
- M. Johannesson, Quality-adjusted life-years versus healthy-years equivalents: a comment, *J. Health Econ.*, 14 (1995) 9–16.

- M. Johannesson, J.S. Pliskin and M.C. Weinstein, A note on QALYs, time tradeoff, and discounting, *Med. Decision Making*, 14 (1994) 188–193.
- R.J. Labelle and J.E. Hurley, Implications of basing health-care resource allocations on cost-utility analysis in the presence of externalities, *J. Health Econ.*, 11 (1992) 259–277.
- J.R. Landis and G.G. Koch, The measurement of observer agreement for categorical data, *Biometrics*, 33 (1977) 159–174.
- G. Loomes, The myth of the HYE, *J. Health Econ.*, 14 (1995) 1–7.
- D.N. McCloskey, The rhetoric of economics, *J. Econ. Lit.*, 21 (1983) 481–517.
- E. Nord, The QALY: a measure of social value rather than individual utility? *Health Econ.*, 3 (1994) 89–93.
- E. Nord, The person trade-off approach to valuing health care programmes, *Med. Decision Making*, 15 (1995) 201–208.
- J.S. Pliskin, D.S. Shepard and M.C. Weinstein, Utility functions for life years and health status, *Operations Research*, 28 (1990) 206–220.
- J.L. Read, R.J. Quinn, D.M. Berwick, H.V. Fineberg and M.C. Weinstein, Preference for health outcomes: comparison of assessment methods, *Med. Decision Making*, 4 (1984) 315–329.
- J. Richardson, Cost utility analysis: what should be measured? *Social Sci. Med.*, 39 (1994) 7–20.
- M. Rutten-van Mölken, C. Bakker, E. van Doorslaer and S.J. van der Linden, Methodological issues of patient utility measurement, *Med. Care* 33 (1995) 922–937.
- A.K. Sen, *Collective Choice and Social Welfare*, Holden-Day, San Francisco, CA.
- P.F.M. Stalmeier, T.G.G. Bezembinder and I.J. Unic, Proportional heuristics in time trade-off and conjoint measurement, *Med. Decision Making* 16 (1996) 36–44.
- G.W. Torrance, Social preferences for health states. An empirical evaluation of three measurement techniques, *Socio-Econ. Plan. Sci.*, 10 (1976) 129–136.
- G.W. Torrance, Measurement of health state utilities for economic appraisal: a review, *J. Health Econ.*, 5 (1986) 1–30.
- G.W. Torrance and D. Feeny, Utilities and quality-adjusted life years, *Int. J. Technol. Assess. Health Care*, 5 (1989) 559–575.
- G.W. Torrance, W.H. Thomas and D.L. Sackett, A utility maximization model for evaluation of health care programmes, *Health Services Res.*, 7 (1972) 118–133.
- A. Tversky and D. Kahneman, Loss aversion in riskless choices: a reference dependent model, *Q. J. Econ.*, 106 (1991) 1039–1061.
- P.P. Wakker, *Additive Representations of Preferences: A New Foundation of Decision Analysis*, Kluwer, Dordrecht, 1989.
- M.C. Weinstein and H.V. Fineberg, *Clinical Decision Analysis*, Saunders, Philadelphia, PA, 1980.
- A. Williams, The value of QALYs, *Health Social Service J.*, 95 (1985) 3–5.
- A. Wolfson, A. Sinclair, C. Bombardier and A. McGeer, Preference measures for functional status in stroke patients. Interrater and intertechnique comparisons, in R.L. Kane and R.A. Kane (eds), *Values and Long Term Care*, Lexington Books, Lexington, MA, 1982.