Correcting Biases in Standard Gamble and Time Tradeoff Utilities

Sylvie M. C. van Osch, MSc, Peter P. Wakker, PhD,
Wilbert B. van den Hout, PhD, Anne M. Stiggelbout, PhD

The standard gamble (SG) method and the time tradeoff (TTO) method are commonly used to measure utilities. However, they are distorted by biases due to loss aversion, scale compatibility, utility curvature for life duration, and probability weighting. This article applies corrections for these biases and provides new data on these biases and their corrections. The SG and TTO utilities of 6 rheumatoid arthritis health states were assessed for 45 healthy respondents. Various corrections of utilities were considered. The uncorrected TTO scores and the corrected (for utility curvature) TTO scores provided similar results. This article provides arguments suggesting that the TTO scores are biased upward rather than having balanced biases. The only downward bias in TTO scores was small and probably cannot offset the upward biases. The TTO scores are higher than the theoretically most preferred correction of the SG, the mixed correction. These findings suggest that uncorrected SG scores, which are higher than TTO scores, are too high. ***Key words:*** *utility assessment; bias; loss aversion; utility curvature; probability weighting.* ***(Med Decis Making 2004;24:511–517)***

Utilities can be used to measure the effects of treatment outcomes, and they play an important role in cost-effectiveness analyses.[1,2] Two methods to measure the utility of health states are the time tradeoff (TTO) method and the standard gamble (SG) method.[3] Based on normative expected-utility arguments, the SG method has often been considered the gold standard for utility measurement. However, there is much empirical evidence demonstrating that expected utility is not descriptively valid and that its violations generate upward biases in SG utilities.[4–6]

Less is known about the effects of biases in the TTO measurements. Some recent articles have suggested that these biases might neutralize each other,[4] so that no systematic overall bias results. It would then follow that, on average, TTO utilities are closer to true utilities than SG utilities are. This would entail a theoretical justification for the preference for the TTO method that is indeed observed in practice. Another justification for this preference is based on the higher face validity of TTO results than of SG results. In the latter, respondents have been commonly found to exhibit overly extreme risk aversion.[7] This article provides new insights into correction methods for the aforementioned biases, advanced in the economic literature, and tests them in the medical domain.

BIASES IN TTO AND SG UTILITIES

Bleichrodt provided an overview of the biases in utility measurement and their likely effects.[4] We discuss these biases below and summarize them in Table 1.

Utility Curvature

The TTO assumes that the utility of life duration is linear.[3,8] This assumption is, in general, not correct.[9] Empirical evidence shows that the utility of life years is

Received 25 November 2003 from the Department of Medical Decision Making (SMCvO, AMS, WBvdH), Leiden University Medical Center, Leiden, the Netherlands, and Creed (PPW), University of Amsterdam, Amsterdam, the Netherlands. The abstract of this article was presented at the 24th annual meeting of the Society for Medical Decision Making in Baltimore, Maryland, on 20 October 2002 and was presented at the Netherlands Forum for Medical Decision Making, Nijmegen, the Netherlands on 4 April 2003. Financial support for this study was provided entirely by a grant from the Netherlands Organization for Health Research and Development–Medical Sciences (ZonMw). The funding agreement ensured the authors' independence in designing the study, interpreting the data, and writing and publishing the report. Revision accepted for publication 6 July 2004.

Address correspondence and reprint requests to Sylvie M. C. van Osch, Leiden University Medical Center, Department of Medical Decision Making, J10-S, PO Box 9600, 2300 RC Leiden, the Netherlands; phone: +31 (0)71-5264570; fax: +31 (0)71-5266838; e-mail: s.m.c.van_osch@lumc.nl.

DOI: 10.1177/0272989X04268955

Table 1 Summary of Biases Discussed and Their Effects per Method

	Utility Curvature	Probability Weighting	Loss Aversion	Scale Compatibility	Total Effect
Time tradeoff	Down	Not applicable	Up	Up	?
Standard gamble	Not applicable	Up (mostly)	Up	Unknown	Up

concave for most people, with nearby years valued more than remote years.[10]

In TTO measurements, respondents are asked to trade future years, which are, thereby, overweighted in the TTO calculations. This leads to a downward bias of the resulting utilities. SG measurements are not distorted by utility curvature for life duration.

Probability Weighting

Probability weighting entails that people process probabilities in a nonlinear manner. The pattern most commonly found is that people tend to overweight small probabilities and underweight large probabilities. The TTO does not use probabilities and hence is not affected by the corresponding biases. Probabilities do play a role in SG measurements, and therefore probability weighting does affect SG utilities.

Empirical studies of probability weighting include those by Abdellaoui,[11] Bleichrodt and Pinto,[12] Gonzalez and Wu,[13] and Tversky and Kahneman.[14] Probabilities of >0.33 are usually underweighted, so that respondents choose excessively high probabilities to generate indifference in SG questions. This leads to an overestimation of utility in SG measurements. Reversed effects occur for probabilities <0.33, leading to an underestimation of utility. Because utilities of health states usually exceed 0.33, probability weighting will usually generate an upward bias for the SG utilities.[4]

Loss Aversion

Loss aversion refers to the finding that people are more sensitive to losses than to gains.[14] Consequently, losses weigh more heavily in decisions than gains do. Whether an outcome is perceived as a gain or a loss depends on the reference point, which is often the status quo. The TTO takes an impaired health state as the starting point. This starting point is a natural candidate to serve as the reference point for the respondents. The TTO asks how many life years a person is willing to give up to regain optimal health. The person is asked to trade off life years (a loss) for optimal health (a gain). Loss aversion will make people more reluctant to give up life years. Consequently, loss aversion generates an upward bias for the TTO, thus overestimating the utility of health states.

In the SG, the gambles can be perceived as yielding all losses, all gains, or as mixed (yielding both gains and losses), depending on the perceived reference point. It has been argued that the health state being evaluated is most likely to be perceived as the reference point,[15,16] which can be seen as follows. In SG measurements, 2 options are considered. Option 1 with certainty yields an intermediate outcome, that is, the health state to be evaluated. Option 2 is a gamble yielding a good outcome with probability P and a bad outcome with probability $1 - P$. The probability P is varied until indifference results. The certain outcome is not varied and is therefore most naturally taken as the reference point.[15,16] In option 2, the good outcome is then perceived as a gain and the bad outcome as a loss. Consequently, it has been argued that the gamble as a whole is perceived as mixed. If so, for a person who is averse to loss, the gain probability P must then be extra high to offset the loss probability $1 - P$. Loss aversion therefore generates an upward bias in SG utilities.

Scale Compatibility

A less well-known bias is scale compatibility. It refers to the finding that the higher the compatibility of a characteristic with the response scale used, the more attention and weight an individual will give to that characteristic.[4,5,17,18] For the TTO, the response scale is the number of years in good health. More attention is therefore given to duration than to health status. A respondent will be less willing to trade off life years, disregarding the health impact for those years. Thus, higher scores result.

For the SG, the response scale is a probability. Thus, respondents will pay more attention to the probabilities. This may hold as well for the good-outcome probability as for the bad-outcome probability.[4] Therefore, no systematic bias for SG utilities can be predicted.

Corrections for Biases

Methods have been proposed to correct TTO utilities of health states for utility curvature for life duration using the certainty equivalent (CE) standard gamble to assess the utility of length of life.[7] Although quantitative corrections of TTO utilities for loss aversion and scale compatibility are highly desirable, no such corrections are known at present, unfortunately. We can, therefore, present only a correction of TTO utilities for utility curvature for life duration. The corresponding formula is given in Appendix A. For SG utilities, corrections for the biases mentioned have been proposed,[19] with the exception of scale compatibility. We consider 3 possible versions, depending on whether the gamble outcomes are perceived as all gains, all losses, or mixed. Figure 1 shows the corrected SG utilities for each possible perception. The corresponding formulas are given in Appendix B.

We examine the convergent validity of the various corrections proposed and the extent to which the biases in TTO measurements neutralize each other. We speculate on which (corrected) measurements yield utilities closest to true utilities.

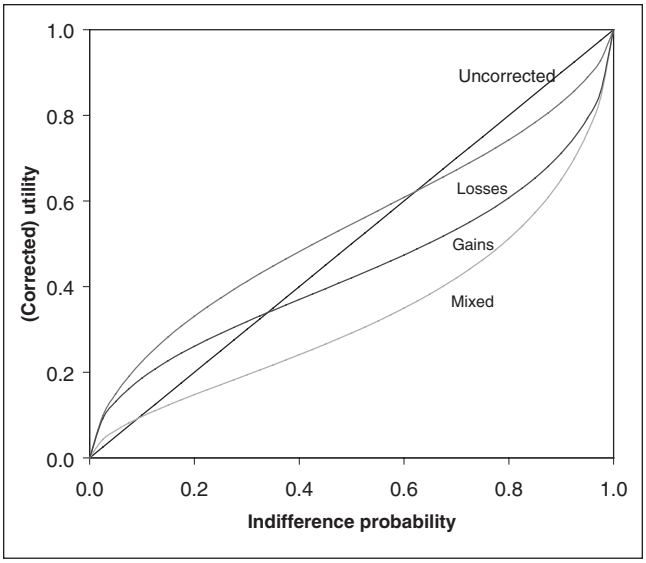

Figure 1 The inverse S-shaped correction functions of standard gamble utilities per perception: all gains, all losses, and mixed. The uncorrected function is also depicted.

METHODS

Procedure

Forty-five respondents were recruited through newspaper ads and pamphlets. They were paid €22.50 for participation. Six rheumatoid arthritis health state descriptions were selected from the descriptions given by rheumatic patients in the Rheumatoid Arthritis Patients in Training Study.[20] Descriptions were taken from the EQ-5D system, a multiattribute health utility system. The EQ-5D system comprises 5 dimensions of health (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression). Each dimension comprises 3 levels (no problems, some/moderate problems, and extreme problems). A unique EQ-5D health state is defined by combining 1 level from each of the 5 dimensions. The health states were chosen so as to cover the utility continuum (0–1), using corresponding EQ-5D valuations based on the TTO.[21] We used the EQ-5D health state descriptions; 21232 (utility of 0.09); 22322 (utility of 0.19); 21322 (utility of 0.36); 21222 (utility of 0.62); 21211 (utility of 0.81); and 21111 (utility of 0.85).

The TTO, SG, and CE were all computerized using the program Ci3.[22] All elicitations were based on the ping-pong search procedure. This procedure leads to fewer inconsistencies in people's preferences than the procedure of direct matching.[23]

All respondents performed 2 sessions with a 2-week interval in between. The order was randomized. Session A consisted of SG and TTO elicitations. The order of elicitations within this session was randomized per method. Session B was devoted to the CE life-year gambles. Each session took 90 min on average to complete and was preceded by oral and written instructions. At any time during an elicitation, it was possible for respondents to take a break, check earlier answers, and possibly change them. At the end of each elicitation, respondents were requested to verify if they indeed considered the two options equivalent.

Session A: SG and TTO

Session A started with a short explanation of rheumatoid arthritis. In total, 6 SGs and 6 TTOs were performed, 1 elicitation for each rheumatoid arthritis health state. In the SG, 2 options were given. Option 1 was a rheumatoid arthritis health state for the respondent's remaining life expectancy (LE). Option 2 was a gamble between good health for LE with probability P and death within a week with probability $1 - P$. Probabilities in the gamble were varied until indifference resulted. LE was based on a respondent's remaining LE derived from Dutch life tables.[24]

For the TTO, respondents were offered the choice between either a rheumatoid health state during LE

and a healthy life for period x ($x \leq$ LE). Period x was varied until indifference resulted.

Session B: CE

Respondents performed 7 CE life-year gambles in good health: CE12.5, CE25, CE37.5, CE50, CE62.5, CE75, and CE87.5. CE is an SG for which probabilities are held constant, in our case at $P = 0.5$. The duration of the certain outcome is varied until indifference results. The CE50 is the number of years that a respondent finds equivalent to a 50-50 gamble between LE and death within a week. CE75 is the number of years equivalent to a 50-50 gamble between the LE and CE50. CE25 is the number of years equivalent to a 50-50 gamble between CE50 and death within a week, and so forth. A detailed discussion of the chained CE measurement method used in this article is available in Verhoef and others.[25] As CE measurements were chained, for example, the CE50 was used to derive the CE75, complete randomization was not possible. The order of elicitations within this session was randomized as much as possible.

The CE values, used to correct the TTO measurements for nonlinearity of utility, were analyzed in the traditional way assuming expected utility. A reanalysis of these data through prospect theory, and the location of a reference point appropriate for such an analysis, is the topic of future research. This article focuses on the novelty of the corrected SG measurements and the comparison of these to traditional measurements.

Data Analysis

The formulas used to calculate utilities from the respondents' choices are explained in Appendices A and B. Discrepancies between methods were assessed for all health states using MANOVA with method as a within-subjects factor to determine convergent validity between the TTO and the SG, both corrected and uncorrected.

RESULTS

Two of the 45 respondents were excluded from the analysis because they were not able to perform CE life-year gambles appropriately, either because the subjective LE was much higher than the LE used ("my grandmother and grandfather are alive and well and both 90 years of age; the 76 years [LE] you offer is far too short") or due to religious arguments ("God decides what will happen, not I"). The respondents consisted of 26 women (mean age = 27, $s = 12$) and 17 men (mean age= 34, $s = 14$). All respondents had received at least a high school education. About 50% of the respondents were university students, and 25% of the respondents had children.

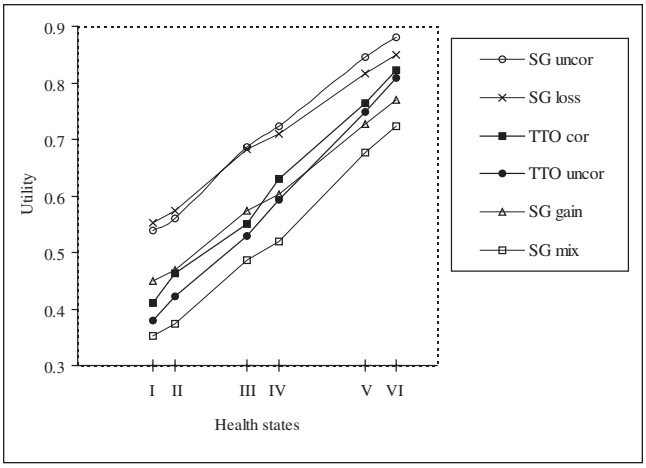

Figure 2 Mean utility for each health state per method and possible corrections.
Note: SG = standard gamble; uncor = uncorrected; TTO = time tradeoff; cor = corrected. The 6 health states are ranked on the x axis according to the corresponding mean utility.

Most respondents (65%) exhibited risk aversion in the CE questions, that is, their CEs were lower than the expected values of the gambles. About 25% of the respondents exhibited risk seeking, and 10% exhibited risk neutrality (the power coefficient r of utility between 0.95 and 1.05). The mean power coefficient r of utility was 1.16 ($s = 1.07$), and the median power coefficient r was 0.80. For the TTO, utility-curvature correction, using the individual r values (corrected TTO), leads to slightly higher scores than uncorrected TTO scores. Figure 2 shows the minor and not significant effect of the correction on the average TTO valuation per health state ($P = 0.29$).

Figure 2 also presents health state utilities as assessed by the SG, both uncorrected and corrected. It shows that uncorrected SG and losses-corrected SG, leading to very similar utilities, always provide the highest value for a health state, followed by gains-corrected SG. Mixed-corrected SG always provides the lowest utility. This order is in line with the differences shown in Figure 1 (see also Appendix B). Gains-corrected SG shows the strongest convergence with both the corrected TTO ($P = 0.51$) and the uncorrected TTO ($P = 0.74$). The losses-corrected SG is relatively high and shows the least convergence with the uncorrected TTO ($P < 0.001$) and the corrected TTO ($P < 0.002$). Mixed-corrected SG provides scores that are

considerably lower than uncorrected TTO scores ($P = 0.05$) or corrected TTO scores ($P < 0.01$).

DISCUSSION

In health economics, the TTO has been developed as an alternative to the SG.[3] Although lacking the theoretical foundations of the SG, the TTO has emerged as the most frequently used method. The main reasons for TTO's wide acceptance are its better feasibility, its higher discriminative power, and its better face validity. The epithet of the SG as gold standard has faded during years of practice. TTO seems to have been accepted as a practical gold standard.

In our data, utility of life years was nearly linear at the aggregate level, and hence correcting the TTO for utility curvature had only a minor effect. Some other studies found stronger deviations from linearity for the utility of life years.[26] Stiggelbout and others used a time frame of 10 years and interviewed disease-free testicular patients who evaluated a good health state, and therefore their findings may not be comparable to ours.[26]

In our data, correcting for utility curvature had no effect. Consequently, this correction did not neutralize the upward bias in TTO due to loss aversion and scale compatibility, resulting in an overall upward bias in TTO scores. This suggests that the even higher uncorrected SG and losses-corrected SG scores are way too high. There is other evidence suggesting that SG scores are too high.[15] No quantitative estimations are known of the effects of loss aversion and scale compatibility on the TTO scores, and hence we cannot estimate the degree of overestimation comprised in TTO scores. In Bleichrodt and Pinto[18] and Bleichrodt and others,[27] similar high durations were used, and no loss aversion was found for such high durations.

The gains-corrected SG showed the strongest convergence with the uncorrected TTO data. However, in our data, the TTO seems to be too high, and thus gains-corrected SG is probably too high also. The mixed-corrected SG may provide better approximations of true utility than the gains-corrected SG. A psychological argument in favor of the mixed-corrected SG is that the certain outcome is fixed in the SG.[19] The framing of the instructions, in which respondents were asked to imagine that the certain health state is their status quo, provides another argument in favor of the mixed correction. Furthermore, immediate death is not plausible to serve as a reference point because it is remote from the actual situation faced by the respondents, which is another reason why it is unlikely that all outcomes in the SG will be perceived as gains. This probably is too distant from a healthy person's status quo, which includes life expectancy. Little is known about the psychology behind the location of the perceived reference point. Qualitative data to provide further insights will be desirable.

CONCLUSION

In our study, utility curvature was absent at the average level, and as a result, correcting TTO scores for utility curvature had little effect at the aggregate level. The loss correction of the SG also had little effect. The gains correction of the SG had more effect, leading to lower scores that were close to the TTO scores and yielding the strongest convergent validity. The mixed-correction of the SG led to considerably lower scores. Besides the convergent validity, Bleichrodt[4] suggested another argument based on conjectured neutralizing biases favoring TTO scores. We have suggested, to the contrary, a net upward bias for TTO scores. There are also theoretical arguments, based on prospect theory, favoring the mixed-correction of the SG. Because we found that TTO scores were higher than mixed-corrected SG scores were, this suggests again that TTO scores are too high, in deviation from what has been thought before. This finding suggests once more that the (even higher) SG scores are much too high.

APPENDIX A
Time Tradeoff Calculations

Estimates for utilities of the 6 health states were derived from the time tradeoff (TTO) questions by dividing the number (x) of years in good health by the life expectancy (LE). A power function with parameter r was used to describe utility of life years. Power functions were chosen because there is empirical evidence supporting these functions.[9] For each respondent, r was estimated and used to correct the respondent's TTO. Following Pliskin and others,[9] the utility function U(Y, Q) for life years Y in health state Q is $U(Y, Q) = bY^rH(Q)$, where H(Q) is a quality-adjusted factor, scaled from 0 to 1. The following argument is taken from Miyamoto and Eraker[7]:

For CEn, n = 25, 50, 75:

$$n/100 = U(CEn, Q)/U(LE, Q).$$

Expanding the right side yields

$$n/100 = bCEn^{r}H(Q)/(bLE^{r}H(Q)) = (CEn/LE)^{r}.$$

Taking logarithms and dividing through yields

$$(1/r)\ln(n/100) = \ln(CEn/LE).$$

A least-squares estimate can be obtained for $(1/r)$.

It can be shown that H(Q), the measure of health quality, is estimated by (x/LE) from the TTO raised to the power *r*.

If a respondent is indifferent between (LE, Q) and (x, Q_{max}), then $(U(LE, Q) = U(x, Q_{max})$:

$$bLE^{r}H(Q) = bx^{r}H(Q_{max}) = bx^{r}, \text{ because } H(Q_{max}) = 1.0.$$

$H(Q) = (x/LE)^{r}$ now follows.[7,26]

APPENDIX B
Standard Gamble Calculations

The following utility calculations are based on prospect theory, following Bleichrodt and others.[19] We use the following notation.

- P = indifference probability provided by the respondent
- U(h) = utility of health state h
- $\omega(P)$ = weight of the probability P
- γ = parameter in the probability weighting function
- λ = loss aversion parameter (value = 2.25)

Tversky and Kahneman proposed the following probability weighting function[14]:

$$\omega(P) = P^{\gamma}/((P^{\gamma} + (1 - P)^{\gamma})^{1/\gamma}).$$

The formula has been found to be different for losses than for gains.

- $\omega^{-}(P)$ = weight of probability of a loss
- $\omega^{+}(P)$ = weight of probability of a gain

If individual estimates of the parameters of the respondent for the relevant outcomes are available, then these values should obviously be used. Such estimations are, however, hard to obtain and are not commonly available in the health literature. In the absence of such information, it seems natural to use the estimations most commonly accepted in the literature, being those by Tversky and Kahneman[14]: $\gamma = 0.69$ for losses and $\gamma = 0.61$ for gains. For a detailed discussion of this point, see section 4 of Bleichrodt and others.[19]

If all outcomes are perceived as gains, then the formula for the standard gamble (SG) utility of the health state is

$$U(h) = \omega^{+}(P).$$

If all outcomes are perceived as losses, then the formula for the SG utility of the health state is

$$U(h) = 1 - \omega^{-}(1 - P).$$

For the mixed case, the formula for the SG utility of the health state is

$$U(h) = \omega^{+}(P)/(\omega^{+}(P) + \lambda\omega^{-}(1 - P)).$$

The SPSS syntax file is available from the authors on request.

REFERENCES

1. Drummond MF, Stoddart GL, Torrance GW. Methods for the Economic Evaluation of Health Care Programmes. 2nd ed. New York: Oxford University Press; 1999.
2. Sox HC, Blatt MA, Higgins MC, Marton KI. Medical Decision Making. Stoneham (MA): Butterworth-Heinemann; 1988.
3. Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. Health Serv Res. 1972;7:118–33.
4. Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Econ. 2002;11(5):447–56.
5. Delquie P. Inconsistent trade-offs between attributes: new evidence in preference assessment biases. Manage Sci. 1993;39(11): 1382–95.
6. Hershey JC, Schoemaker PJH. Probability versus certainty equivalence methods in utility measurement: are they equivalent? Manage Sci. 1985;31:1213–31.
7. Miyamoto J, Eraker SA. Parameter estimates for a QALY utility model. Med Decis Making. 1985;5(2):191–213.
8. Stiggelbout AM, Kiebert GM, Kievit J, Leer JWH, Habbema JDF, de Haes JCJM. The utility of the time trade-off method in cancer patients: feasibility and proportional trade-off. J Clin Epidemiol. 1995;48(10):1207–14.
9. Pliskin JS, Shepard DS, Weinstein MC. Utility functions for life years. Health Status Utility Functions. 1979;207–24.
10. Bleichrodt H, Pinto JL. The validity of QALY's under non-expected utility. Economic J. In press.
11. Abdellaoui M. Parameter-free elicitation of utility and probability weighting functions. Manage Sci. 2000;46(11):1497–512.
12. Bleichrodt H, Pinto JL. A parameter-free elicitation of the probability weighting function. Manage Sci. 2000;46(11):1485–96.
13. Gonzalez R, Wu G. On the shape of the probability weighting function. Cognit Psychol. 1999;38(1):129–66.
14. Tversky A, Kahneman D. Advances in prospect theory: cumulative representation of uncertainty. J Risk Uncertain. 1992;5:297–323.

15. Hershey JC, Schoemaker PJH. Probability versus certainty equivalence methods in utility measurement: are they equivalent? Manage Sci. 1985;31:1213–31.
16. Robinson A, Loomes G, Jones-Lee M. Visual analog scales, standard gambles, and relative risk aversion. Med Decis Making. 2001; 21:17–27.
17. Tversky A, Sattath S, Slovic P. Contingent weighting in judgment and choice. Psychol Rev. 1988;95(3):371–84.
18. Bleichrodt H, Pinto JL. Loss aversion and scale compatibility in two-attribute trade-offs. J Math Psychol. 2002;46(3):315–37.
19. Bleichrodt H, Pinto JL, Wakker PP. Making descriptive use of prospect theory to improve the prescriptive use of expected utility. Manage Sci. 2001;47(11):1498–514.
20. de Jong Z, Munneke M, Zwinderman AH, et al. Is a long-term high-intensity safe in patients with exercise program effective and rheumatoid arthritis? Results of a randomized controlled trial. Arthritis Rheum. 2003;48(9):2415–24.
21. Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997;35:1095–108.
22. Ci3-250. USA: Skim Software. Rotterdam, the Netherlands: Sawtooth Software, Inc.; 2000.
23. Lenert LA, Cher DJ, Goldstein MK, Bergen MR, Garber A. The effect of search procedures on utility elicitations. Med Decis Making. 1998;18(1):76–83.
24. Centraal Bureau voor de Statistiek. Overlevingstafels 2001. 1 June 2003. Available from: http://www.cbs-nl.
25. Verhoef LCG, de Haan AFD, van Daal WAJ. Risk attitude in gambles with years of life: empirical support for prospect theory. Med Decis Making. 1994;14:194–200.
26. Stiggelbout AM, Kiebert GM, Kievit J, Leer JWH, Stoter G, de Haes JCJM. Utility assessment in cancer patients: adjustment of time trade-off scores for the utility of life years and comparison with standard gamble scores. Med Decis Making. 1994;14:82–90.
27. Bleichrodt H, Pinto JL, Abellan JM. A consistency test of the time trade-off. J Health Econ. 2003;22:1037–52.