The construction of health state utilities

Osch, S.M.C. van

ISBN: 978-90-9021909-7

Printed by: Grafisch Bedrijf Ponsen & Looijen b.v.

Cover design: M. Broere & S.M.C. van Osch

The Construction of Health State Utilities

Proefschrift

ter verkrijging van
de graad doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. P.F. van der Heijden,
volgens besluit van het College van Promoties te verdedigen
op donderdag 6 september 2007
klokke 15.00 uur

door

Sylvie Michaela Cornélie van Osch
geboren te Bladel
in 1976

Promotiecommissie

Promotores:	Prof. dr. A.M. Stiggelbout
	Prof. dr. P.P. Wakker
	Erasmus Universiteit Rotterdam
	Universiteit Maastricht
	Prof. dr. J. Kievit
Referent:	Prof. dr. H. Bleichrodt
	Erasmus Universiteit Rotterdam
Lid:	Dr. W.B. van den Hout

The studies described in this thesis were performed at the Department of Medical Decision Making of the Leiden University Medical Center, Leiden, The Netherlands, and were financially supported by a grant (number 904-66-089) from The Netherlands Organization for Health Research and Development – Medical Sciences (ZonMw).

The financial support of the J.E. Jurriaanse Stichting in printing this thesis is gratefully acknowledged.

Voor Bart

Contents

1

Introduction

Weather forecasting is the science of predicting the state of the atmosphere for a future time and location. However, any model cannot include all variables that are relevant to it, e.g. the weather prediction cannot incorporate the wind caused by the wings of a butterfly.

When a decision is to be made, there are, by definition, two or more possible actions. Each action leads in one or more possible outcomes. Some outcomes will be more preferred than others, and some will be more likely to occur. A decision can be based on emotions, it can be taken intuitively or it can be a reasoned decision. A decision maker may determine the likelihood of each outcome to occur and attach a value to that outcome, and make a (possibly rational) decision. In medical decision-making different treatments (actions) can lead to different outcomes certain risks. Utilities can be elicited to measure the health benefits of treatments, in other words, be used to 'attach values to an outcome'. Hence, the measurement of utility is central in medical decision-making. However, the utility measurement methods themselves are fraught with inconsistencies and biases. Payne et al. argue that preferences, and thus utilities are constructed during elicitation rather than elicitation being a form of uncovering existing values (1). This thesis deals with the measurement of health state utilities. The aim of this thesis is to study decision-making from both a normative (prescriptive) and an empirical (descriptive) point of view. Purpose is to improve the measurement of utility in health care through findings from non-expected utility theory, i.e. cumulative prospect theory (PT).

"Economists often criticize psychological research for its propensity to generate lists of errors and biases, and for its failure to offer a coherent alternative to a rational-agent model … this is just another way of saying that rational models are psychologically unrealistic" wrote Kahneman (2). Psychology indeed provides a description of how people make judgments and decisions, and thus proposes a descriptive decision-

making model. A normative theory specifies an optimal set of decision rules. Such as, if medical treatment A is preferred over treatment B, and treatment B is preferred over treatment C; it follows that treatment A is to be preferred over treatment C. A descriptive model tries to represent behavior or the anticipation of behavior. The goal is then to obtain an accurate model of the existing decision process. If discrepancies exist between normative theory and observed behavior, a further question is what people should be told to do in order to satisfy certain goals, i.e. what should be prescribed to them. For this we need a prescriptive theory. Prescriptive models are based on normative theories. Thus, descriptive and prescriptive decision-making models differ in how the parameters comprising the models are obtained. Prescriptive and descriptive decision making models bring forward different perspectives. The descriptive decision making model can be evaluated afterwards by assessing the validity of the model through the reproduction of the behavior of the decision maker. A prescriptive model is to be evaluated according to the ability to bring about decisions that are optimal, and that lead to the predicted (desirable) outcomes.

In medical decision-making, the most commonly used decision making theory is expected utility (EU) theory. EU is based on the axioms (i.e. normative decision rules) formulated by von Neumann and Morgenstern (3). To apply EU, it is essential to quantify uncertainties in terms of probabilities, and values of outcomes in terms of utilities. The expected utility of each available action, obtained by combining probabilities and utilities of its associated outcomes, is used to determine the optimal action (4). Empirical evidence of biases in utility measurement is well documented in management science, economics and psychology (5-7). In medicine, however, the awareness of these biases is much more restricted.

Based on normative EU arguments, the standard gamble (SG) method has historically been considered the gold standard for utility measurement. The major violations of EU are explained by Prospect Theory (PT), a descriptive decision making theory (5;11). These violations include loss aversion and probability weighting. However there is ample empirical evidence that EU is not descriptively valid, and that its violations generate upward biases in SG utilities (8-10). Loss aversion refers to the finding that people are more sensitive to losses than to gains. This, for instance, explains why a casino player at the end of the evening may take high risks to work away his losses. Or why a stockholder experiences more difficulties in selling stocks that have decreased in value than stocks that have increased in value. Probability weighting entails that people process probabilities in a nonlinear manner. The pattern that is most often found is that people overweight small probabilities and underweight large probabilities.

The SG generally requires a respondent to compare the certainty of being in the health state to be valued for the remaining life expectancy, with a gamble that offers a chance (probability p) of optimal health for the remaining life expectancy but also entails a risk of immediate death (probability 1-p). In the generally used probability equivalent of the SG, the probability p is varied so as to identify the point at which the respondent is indifferent to the choice between the health state and the gamble. The utility of the health state is calculated by equating the expected utilities of the two alternatives. In health economics, the time trade-off (TTO) has been developed as an alternative to the SG (10). In the TTO, the subject is asked how many years in optimal health she/he considers to be equivalent to a period (e.g. their remaining life expectancy) in a particular impaired health state. A major and basic difference with the SG is that the TTO is not based on expected utility – in the sense of being a product of probability and outcome - and thus provides a riskless measure. TTO utilities have been found to

better reflect individual preferences for health than SG utilities do (6), and have a higher face validity. The most well known biases to affect TTO utilities are scale compatibility, utility curvature and loss aversion (6). Unfortunately, these prominent biases in the TTO measurements have rarely been investigated.

With respect to societal decision-making, utility is included in cost utility analysis. In which the benefits of health care programs are expressed in utility terms and are compared to costs. The use of biased utilities will lead to biased resource allocation decisions. Therefore the joint effect of the aforementioned biases should be minimized in the elicitation of utilities, whether using the SG or using the TTO. This requirement creates a need for more knowledge about these biases in health utility measurement, so as to adequately correct for them or at least to be aware of their effect on utilities (4).

The VAS is not a preference-based measure, and therefore does not provide utilities. It is nevertheless often substituted for SG or TTO, for reasons of feasibility. Therefore, we will evaluate this measure as well. In medical decision making, until now mostly standard gambles have been used to assess risk attitude. Elicitation is a complex task, fraught with biases. We aim to develop the health-risk attitude scale (HRAS) in order to assess health risk attitude.

The present thesis is based on 9 chapters. In the second chapter the SG and TTO are introduced as methods to assess utilities. The potential biases in these methods that are discussed are loss aversion, probability weighting, scale compatibility, and utility curvature for life duration. This chapter describes correction methods for the aforementioned biases that have been advanced in the economic literature, and tests them in the medical domain. No clear conclusions can be drawn yet in this chapter, because information is lacking on some crucial premises regarding the corrections.

Such information is the topic of subsequent chapters. We then explore, with the use of qualitative research, how utilities are constructed, and which themes and biases influence utilities.

Chapter 3 combines qualitative with quantitative data, so as to provide evidence of the reference point in life-year certainty equivalent (CE) standard gambles and to explore the psychological basis of the reference point. Risk behavior has been shown to depend strongly on the perception of the outcome as either a gain or a loss. According to prospect theory, the reference point, i.e. a point of view, determines how an outcome is perceived. However, no theory on the location of the reference point exists for a standard gambles (whether probability equivalent or certainty equivalent). Additionally, for the health domain, there is no direct evidence for the location of the reference point. With knowledge of the reference point, the proper correction method can be applied to counteract the biases probability weighting and loss aversion.

Chapters 4 and 5 also contain a combination of qualitative and quantitative data, but now for the probability-equivalent SG and for the TTO. The effect of the aforementioned biases on SG and TTO utilities is assessed with the use of qualitative data. The first objective was to locate the SG outcome that is the reference point or that seems to lie closest to the reference point. The second objective was to obtain an indication of whether respondents focus more on the bad outcome or on the good outcome, in order to assess whether scale compatibility results in a systematic bias upwards or downwards. To assess this point, we determined the focus of attention. Additionally, we aimed to verify that a main focus on a bad outcome or good outcome will lead to higher or lower utilities, respectively. Relevant themes that were raised by respondents and that could result in a biased utility were also taken into account. Chapter 5 provides similar research for the TTO.

A frequently used method to assess valuations is the visual analog scale (VAS). In chapter 6, we explored approaches to valuing a health state on a VAS. Cognitive processes involved in for instance valuing a health state, can be carried out at two distinct levels, each with qualitatively different mechanisms. Dual-processing theory states that thoughts, behaviors and feelings result from the interplay of automatic (and implicit) and controlled (and explicit) processing (12). We carried out two experiments in which respondents were probed for approaches used in the VAS. Possible approaches were explored in the first experiment, and were systematically examined in the second.

In medicine, as well as in medical decision making, risk and uncertainty are almost always standard ingredients. Consequently, risk attitude is highly relevant to medical decision-making. It is therefore surprising, that risk attitude, is generally not, or only implicitly, taken into account. Differences among individuals, whether patient or doctor, in risk attitude and, consequently, in their response to risky medical situations can provide valuable information with respect to (the understanding of) treatment preferences and clinical decisions. The CE gambles described in chapter 3 provide a formal approach to assessing risk attitude according to EU. No alternative to this formal approach, which is cognitively difficult and time-consuming, was available. We therefore decided to develop a health risk attitude scale (H-RAS), which we introduce in chapter 7. The H-RAS aims to assess how persons value their health and manage health risks. The H-RAS is psychometrically tested.

In Chapter 9, the main findings and their implications for health care will be summarized and discussed.

2

Correcting biases in standard gamble and time trade-off utilities

Correcting biases in standard gamble and time trade-off utilities.

S.M.C. van Osch, P.P. Wakker, W.B. van den Hout, A.M. Stiggelbout

Medical Decision Making 2004; 24(5): 511-517.

Abstract

The standard gamble (SG) method and the time trade-off (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 chapter applies corrections for these biases, proposed in the economic literature, and provides new data on these biases and their corrections. The SG and TTO utilities of six 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. The gains-corrected SG showed the best convergence with TTO scores. It has been suggested that TTO biases neutralize each other (whereas SG biases do not), so that the TTO method provides good estimates of utility. This chapter provides arguments suggesting that the TTO scores are biased upwards, rather than having balanced biases. First, the only downward bias in TTO scores (due to utility curvature of life duration) was small and, probably, cannot offset the upward biases. Second, the TTO scores are higher than the theoretically most preferred correction of the SG, the mixed correction. These findings suggest once more that uncorrected SG scores, which are higher than TTO scores, are too high.

Introduction

Utilities can be used to measure the effects of treatment outcomes, and play an important role in cost effectiveness analyses (8;9). Two methods to measure the utility of health states are the time trade-off (TTO) method and the standard gamble (SG) method (10). 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 (6;7;13).

Less is known about the effects of biases in the TTO measurements. Some recent papers have suggested that these biases might neutralize each other (6), 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 (14). This chapter 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 (6). We discuss these biases below, and summarize them in Table 2.1.

Utility curvature

The TTO assumes that the utility of life duration is linear (10;15). This assumption is, in general, not correct (16). Empirical evidence shows that the utility of life years is concave for most people, with nearby years valued more than remote years (17). 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 Abdellaoui, Bleichrodt & Pinto (18), Bleichrodt & Pinto (19), Gonzalez & Wu (20), and Tversky & Kahneman (11). Probabilities $p > .33$ 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 $p< .33$, leading to an underestimation of utility. Because utilities of health states usually exceed .33, probability weighting will usually generate an upward bias for the SG utilities (6).

Loss aversion

Loss aversion refers to the finding that people are more sensitive to losses than to gains (11). 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 in order 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 (7;21), which can be seen as follows. In SG measurements, two options are considered. Option 1 with certainty yields an intermediate outcome, i.e. 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 (7;21). 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 loss averse, 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 (6;13;22;23). 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 (6). Therefore, no systematic bias for SG utilities can be predicted.

Table 2.1. Summary of biases discussed and their effects per method.

	Utility curvature	***Probability weighting***	***Loss aversion***	***Scale compatibility***	***Total effect***
TTO	Down	Not applicable	Up	Up	?
SG	Not applicable	Up (mostly)	Up	Unknown	Up

Corrections for biases

Methods have been proposed to correct TTO utilities of health states for utility curvature for life duration using the Certainty Equivalent standard gamble (CE) to assess the utility of length of life (14). 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, only present a correction of TTO utilities for utility curvature for life duration. The corresponding formula is given in Appendix 2A. For SG utilities, corrections for the biases mentioned have been proposed (4) with the exception of scale compatibility. We consider three possible versions, depending on whether the gamble outcomes are perceived as all

gains, all losses, or mixed. Figure 2.1 shows the corrected SG utilities for each possible perception. The corresponding formulas are given in Appendix 2B.

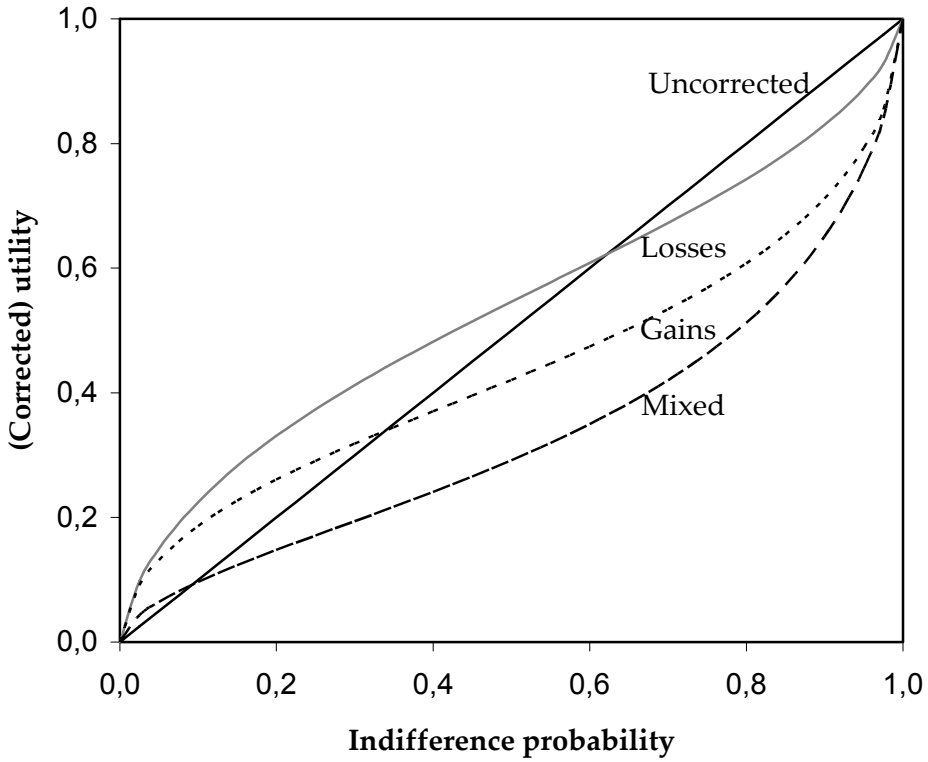

FIGURE 2.1. The inverse S-shaped correction functions of SG utilities per perception: all gains, all losses, and mixed. The uncorrected function is also depicted.

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.

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 (24). Descriptions were taken from the EQ-5D system, a multi-attribute health utility system. The EQ-5D system comprises five dimensions of health (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression). Each dimension comprises three levels (no problems, some/moderate problems, and extreme problems). A unique EQ-5D health state is defined by combining one level from each of the five dimensions. The health states were chosen so as to cover the utility continuum (0–1), using corresponding EQ-5D valuations based on the TTO (25). We used the EQ-5D health state descriptions; 21232 (utility of .09), 22322 (utility of .19), 21321 (utility of .36), 21222 (utility of .62), 21211 (utility of .81), and 21111 (utility of .85).

The TTO, SG, and CE were all computerized using the program Ci3 (26). 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 (27).

All respondents performed two sessions with a two-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 minutes 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, standard gamble and time trade-off

Session A started with a short explanation of rheumatoid arthritis. In total, six SG's and six TTO's were performed, one elicitation for each rheumatoid arthritis health state. In the SG, two 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 life expectancy derived from Dutch life tables (28).

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, certainty equivalent (CE)

Respondents performed seven Certainty Equivalent life-year gambles in good health, CE12.5, CE25, CE37.5, CE50, CE62.5, CE75, and CE87.5. CE is a standard gamble for which probabilities are held constant, in our case at p = .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; etc. A detailed discussion of the chained CE measurement method

used in this chapter is available in Verhoef et al. (29). As CE measurements were chained, e.g. the CE50 was used to derive the CE75, complete randomisation 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 chapter 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 2A and 2B. 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 life expectancy 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 females (mean age = 27, s.d. = 12) and 17 males (mean age = 34, s.d. = 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.
Most respondents (65%) exhibited risk aversion in the CE questions, i.e. their CE's 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.d. = 1.07) and the median power coefficient *r* was .80. For the TTO, utility-curvature correction, using the individual *r*-values (corrected TTO), leads to slightly higher scores than uncorrected TTO scores. Figure 2.2 shows the minor and non-significant effect of the correction on the average TTO valuation per health state ($p = .29$).

Figure 2.2 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 2.1 (see also Appendix 2B). Gains-corrected SG shows the strongest convergence with both the corrected TTO ($p = .51$) and the uncorrected TTO ($p = .74$). The losses-corrected SG is relatively high and shows the least convergence with the uncorrected TTO ($p < .001$) and the corrected TTO ($p < .002$). Mixed-corrected SG provides scores that are considerably lower than uncorrected TTO scores ($p = .05$) or corrected TTO scores ($p < .01$).

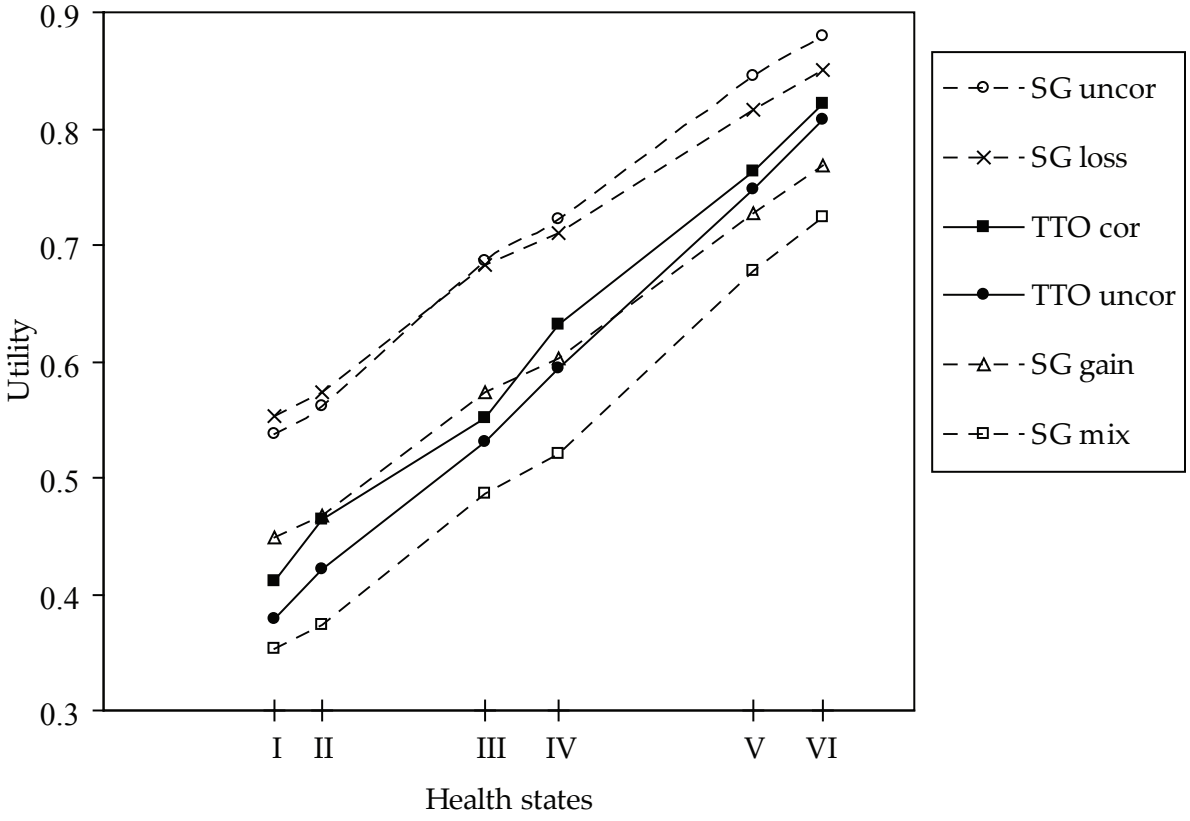

FIGURE 2.2. Mean utility for each health state per method and possible corrections. The six health states are ranked on the x-axis according to the corresponding mean utility.

Discussion

In health economics, the TTO has been developed as an alternative to the SG (10). 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 (30). Stiggelbout et al. used a time frame of ten years and interviewed disease-free testicular patients who evaluated a good health state and, therefore, their findings may not be comparable to ours (30).

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 (7;13). 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 (23) and Bleichrodt, Pinto, and Abellan (31), 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 (4). The framing of the instructions, where respondents were asked to imagine that the certain health state is their status quo, provides another argument in favor of the mixed correction. Further, 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.

Conclusions

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 (2002) 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, 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 2A. TTO calculations

Estimates for utilities of the six health states were derived from the TTO questions by dividing the number (x) of years in good health by the 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 (16). For each respondent, *r* was estimated and used to correct the respondent's TTO. Following Pliskin et al. (16), the utility function U(Y,Q) for life years Y in health state Q is $U(Y,Q) = bY^{r}H(Q)$, where H(Q) is a quality adjusted factor, scaled from 0 to 1. The following argument is taken from Miyamoto and Eraker (14):

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 $U(LE,Q) = U(x,Q_{max})$:

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

$H(Q) = (x / LE)^{r}$ now follows (14;30).

Appendix 2B. SG calculations

The following utility calculations are based on prospect theory, following Bleichrodt, Pinto, & Wakker, 2001 (4). 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 (11):

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

The formula has been found to be different for losses than for gains, in which $\omega^{-}(p)$ = weight of probability of a loss, and $\omega^{+}(p)$ = 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 (11): $\gamma = 0.69$ for losses and $\gamma = 0.61$ for gains. For a detailed discussion of this point see section 4 of Bleichrodt, Pinto, and Wakker (4).

If all outcomes are perceived as gains, then the formula for the 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:

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

3

Exploring the reference point in prospect theory

Gambles for length of life

Exploring the reference point in prospect theory: Gambles for length of life.

S.M.C. van Osch, W.B. van den Hout, A.M. Stiggelbout

Medical Decision Making: 2006; 26: 338-347.

Abstract

Attitude towards risk is an important factor determining patient preferences. Risk behavior has been shown to be strongly dependent on the perception of the outcome as either a gain or a loss. According to prospect theory, the reference point determines how an outcome is perceived. However, no theory on the location of the reference point exists and for the health domain, there is no direct evidence for the location of the reference point. This chapter combines qualitative with quantitative data, to provide evidence of the reference point in life-year CE gambles and to explore the psychology behind the reference point. We argue that goals (aspirations) in life influence the reference point. While thinking aloud, 45 healthy respondents gave certainty equivalents for life-year CE gambles with long and short durations of survival. Contrary to suggestions from the literature, qualitative data demonstrated that the offered certainty equivalent most frequently served as the reference point. Thus, respondents perceived life-year CE gambles as mixed. Framing of the question and goals set in life proved to be important factors behind the psychology of the reference point. On the basis of our quantitative and qualitative data, we argue that goals alter the perception of outcomes as described by prospect theory by influencing the reference point. This relationship is more apparent for the near future as opposed to the remote future, as goals are mostly set for the near future.

Introduction

Utility elicitations based on the normative theory of expected utility are not descriptively valid and the major violations are explained by prospect theory (11). An important point in prospect theory is the location of the reference point, i.e. point of view. According to prospect theory, risk behavior varies depending on whether outcomes are perceived as gains or losses, relative to this reference point. Characteristically, a relative gain accompanied by a risk evokes risk-averse behavior, whereas a relative loss accompanied by a risk evokes risk-seeking behavior. The reference point determines this labeling of an outcome (5) and therefore, is an important factor in explaining risk behavior. Little is known about the psychology behind the location of the reference point, for which prospect theory does not include a hypothesis. This chapter attempts to explore the reference point. In prospect theory, outcomes are expressed as gains or losses (positive or negative deviations) relative to a (neutral) reference point (32). With respect to the health domain, it is observed that risk attitude, and thus the reference point, influences treatment choices made by patients (30;33). Furthermore, standard gamble (SG) utilities are often used in health care and two well-documented biases from expected utility (probability weighting and loss aversion) should be corrected for in standard gamble utility calculations (4;34). This correction requires knowledge of the reference point. Therefore, the absence of a theory on the location of the reference point poses a problem for economic evaluation (4;35).

The only available measurement technique in the health domain to assess risk attitude as described in prospect theory is a SG (29). The probability equivalent is a SG in which probabilities are varied and the certain outcome is held constant. The certainty equivalent (CE) is a SG in which probabilities are held constant and the certain

outcome is varied. In the SG, a person behaves in a risk-seeking way if a risky prospect is preferred to a risk-free prospect of equal or greater expected value. A person behaves in a risk-averse way if a risk-free prospect is preferred to a risky prospect of equal or greater expected value. We assume that the CE life-year gamble is subject to deviations from expected utility as described in prospect theory.

In monetary CE gambles, the status quo often serves as the reference point (4). In health settings, e.g. life-year CE gambles, there is only indirect empirical evidence concerning the reference point, with the following argument: for monetary CE gambles, risk-averse behavior is predominantly observed if the outcomes are perceived as gains. Empirical studies report risk-averse behavior for life-year CE gambles (36). Thus, life-year gambles seem to be processed as gains (4), suggesting that the reference point in life-year CE gambles has been zero life years. Death as a reference point seems counter-intuitive, it seems more psychologically plausible that a respondent's life expectancy influences his/her perception of an outcome as a gain or a loss (29), this would contradict the reasoning that life-year CE gambles are processed as only gains. This would fit Kahneman and Tversky's argument that for some situations the aspiration level may determine if outcomes are perceived as a gain or a loss (5). An aspiration level in CE life-year gambles could be the number of years a person strives for, in order to realize a specific goal. Verhoef et al. observed that respondents accepted more risk in order to achieve their aspiration level. Unfortunately, the latter authors did not assess the motivations of respondents systematically (29;37). The impact of goals and immediate needs on decision making and risk behavior has also been emphasized by Lopes, and Schneider & Lopes (38;39).

The purpose of the present article is to provide more insight into the reference point, by combining qualitative and quantitative data. We consider the three outcomes of the

CE life-year gamble (high outcome of the gamble, low outcome of the gamble, and the certain outcome) as potential reference points. For the rest of the paper we will speak of "an outcome that serves as reference point", meaning that an outcome is closest to, or seems to include, the reference point referred to in prospect theory. To provide indirect evidence on the reference point, quantitative data was gathered on the preference curve for life years. Prospect theory theorizes the inflection point of this preference curve, where the curve alters from convex to concave, to be the reference point (11). Additionally, more direct evidence was obtained with the use of qualitative data, thus also providing information about the psychology behind the reference point. We hypothesized that goals drive the reference point. Qualitative data may reveal the effect of goals on the reference point. Consequently, we reasoned that more attention is paid to the life years during which goals are to be realized. Therefore, assessing the focus of attention could provide (additional) insight into the outcome that serves as reference point and the hypothesized relationship between a reference point and goals.

Methods

Procedure

Forty-five respondents were recruited using newspaper ads and pamphlets. They were paid € 22.50 for participation in two interviews, one of which is the topic of this paper.

Quantitative data

Respondents performed seven life-year CE gambles. Probabilities were held constant at p = .5, and the certain outcome was varied until indifference resulted. We preferred

50-50 gambles for they are cognitively easier, and the probability weighting bias is thought not to have a large effect (4). CE50 is the number of years a person finds equivalent to a 50-50 gamble, between the remaining life expectancy (LE) as a high outcome and death within a week as a low outcome. LE was based on a respondent's remaining life expectancy derived from Dutch life tables (28). To elicit the CE25, the high outcome of the gamble was the previously elicited CE50, and death within a week was the low outcome. Consequently, CE25 is the number of years a respondent finds equivalent to a 50-50 gamble between CE50 and death within a week. Similarly, the CE75 is the number of years a respondent finds equivalent to a 50-50 gamble between the LE as a high outcome and CE50 as a low outcome, etc. The CE50, serving as an outcome for other gambles, had to be performed first, after which the CE25 or the CE75 followed randomly, and thereafter the remaining gambles were performed in a random order. Respondents performed the CE12.5, CE25, CE37.5, CE50, CE62.5, CE75, and CE87.5 (see Figure 3.1).

Elicitations were computerized using the program Ci3 Version 2.5 (26) (Appendix 3A). Certainty equivalents were obtained via a choice-bracketing approach (series of ping-pong questions) that involved forced choices. Each interview took an average of 90 minutes to complete. Respondents started with a verbal and a written explanation, followed by two examples. At any time during an elicitation, it was possible for respondents to take a break, or check earlier answers within that elicitation and change these if they wanted to. Elicitations ended when respondents indicated that they valued two options equally.

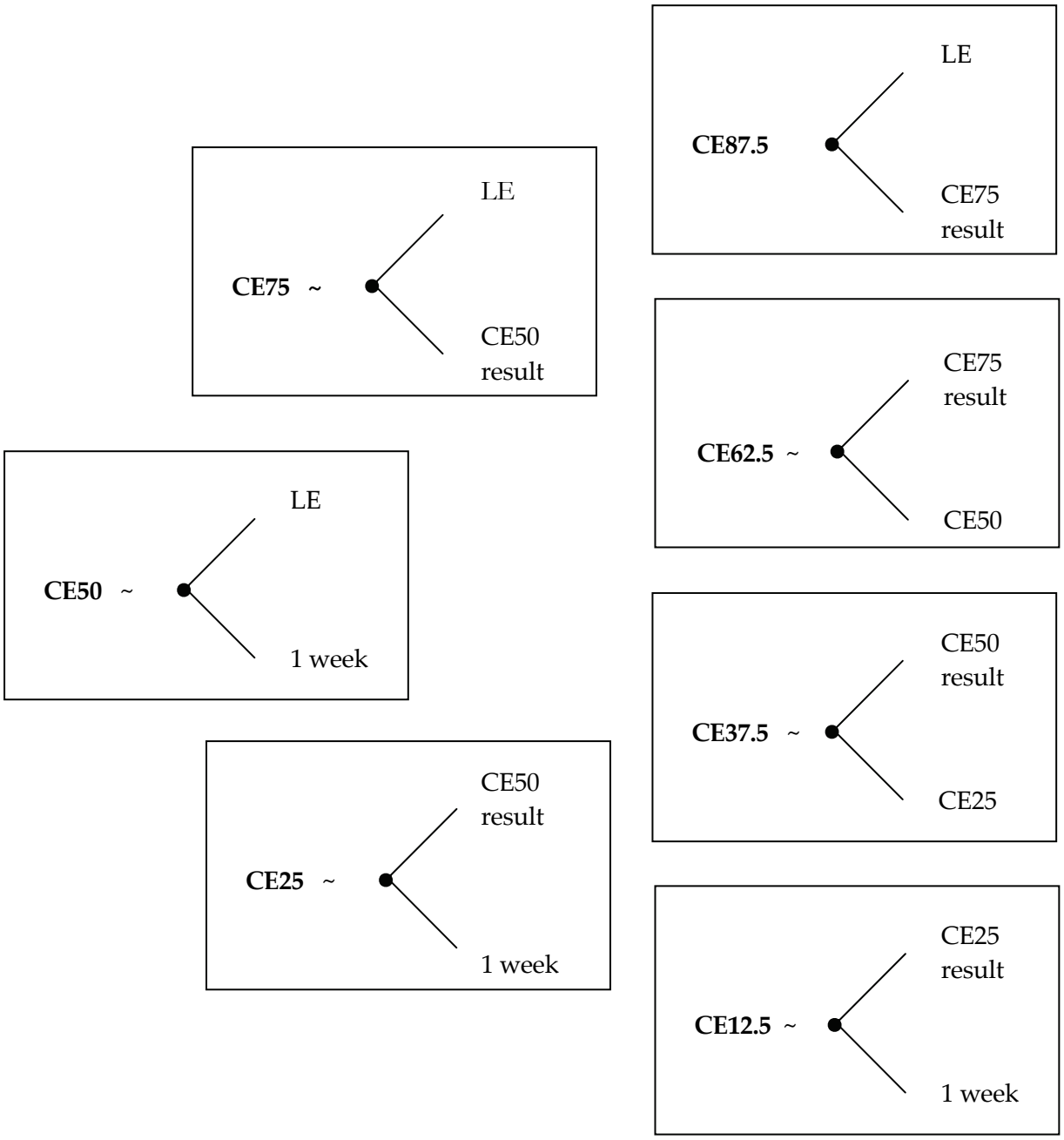

FIGURE 3.1. CE method. The point of indifference in the first gamble, CE50 result, is used to construct the CE25 or CE75 gamble, as high and low outcome of these gambles respectively. In this way, the CE12.5, CE25, CE37.5, CE50, CE62.5, CE75 and CE87.5 were elicited. LE means remaining life expectancy.

Qualitative data

All respondents were instructed to think aloud during the interview. They were specifically instructed not to formulate or make perfect sentences, but merely to think aloud whilst considering the choices. Thinking aloud, i.e. verbalization, has been reported to have an effect only on the time needed to complete the task (40). The main purpose of this study was to assess which outcome in the gamble served as the reference point in CE gambles, while maximizing the degree to which the data could be generalized. Therefore, the number of interventions by the interviewer during computer interviews was minimized, to avoid respondents guessing the construct of interest. The interview was tape-recorded and transcribed. Afterwards, the verbal reports were coded by means of a coding scheme using the computer program QSR NVivo 1.3 (41). Two independent coders each coded all reports, after which, disagreements in coding were discussed to form a consensus coding. If no consensus was reached, a third person was consulted.

Coding

"Reference point" was coded if a point of view was formulated, i.e. respondents used an outcome of the life-year gamble as a starting point to indicate or calculate the difference with another outcome or both other outcomes and, thus, indicating the perception of the outcomes as loss or gain. A coded "reference point" was the outcome in the life-year gamble that seemed closest to, or included, the reference point. In the qualitative analysis, the three outcomes that could serve as the reference point were the high outcome of the gamble, the low outcome of the gamble, and the offered CE. The latter is the certain outcome that was offered in the search procedure.

We coded a high outcome as "reference point" if one or both other outcomes (low outcome and offered CE) were labeled as losses relative to the high outcome. We

coded a low outcome as "reference point" if one or both other outcomes (high outcome and offered CE) were labeled as gains relative to the low outcome.

We coded an offered CE as "reference point" if the high outcome was labeled as a gain and/or the low outcome was labeled as a loss, relative to the offered CE. Table 3.1 shows three possible verbalized remarks and the appropriate codings. The life-year CE gamble evaluated in this example is a gamble with a 50% chance of living for 40 years and a 50% chance of dying within a week (1 week, 0.5, 40 years, option 1) versus the offered CE of 16 years (option 2).

Table 3.1. Example of remarks and codings of the reference points

Example remarks:	**Reference point**
"I can gain 40 years if the gamble goes well or gain 16 years if I choose option 2."	Low outcome (0 years)
"I can gain 24 years if the gamble goes well or lose 16 if it doesn't."	Offered CE (16 years)
"If I choose option 2, I will lose 24 years. If the gamble turns out badly, I will lose all 40 years."	High outcome (40 years)

A "focus of attention" was coded when a respondent mentioned an outcome but did not explicitly compare it to another outcome. For this coding, less strict coding rules applied than for the reference point coding. A "focus of attention" was coded when an outcome was compared to another outcome independently of a reference point being deducible or not. We assumed that the more frequently an outcome was mentioned or compared, the more attention a respondent paid to that outcome.

A "goal" was coded if a statement was made regarding the realization of a goal with respect to an outcome. In other words, if the respondent referred to a survival period stated in an outcome and assessed that period to be too short, or long enough, to realize the goal.

Data analysis

The high outcome of the CE50 was each individual's personal remaining life expectancy (LE) and therefore, the raw certainty equivalent alone does not provide information about risk behavior. To allow for comparison of risk behavior between CEs, the proportional match (PM) was calculated as follows (42). Each certainty equivalent was normalized to the range of the gamble, i.e.:

$$PM = (CE - low) / (high - low).$$

Thus, risk-neutral behavior is indicated by a PM value of 0.5, resulting when the raw CE was chosen equal to the expected value of the gamble. If the CE is chosen higher than the expected value, PM is higher than 0.5, thus denoting risk-seeking behavior. Finally, PM is lower than 0.5 if the raw CE is lower than the expected value, i.e. the respondent displays risk-averse behavior.

Additionally, for each respondent separately, a logistic utility curve was fitted to the seven raw CE values[i]. The logistic curve was argued by Kahneman and Tversky (11) and empirically tested by Verhoef et al. (9):

$$U(t) = \alpha / (1 + (\beta / t)^{\gamma}).$$

Parameter α was chosen so that U(LE) = 1 and parameters β and γ were numerically set to minimize the squared differences between the CE and U^{-1} values. Goodness of fit was measured by the explained variance (R^2). The logistic utility curve is s-shaped and changes from convex to concave at the inflection point:

$$t^* = \beta\,((\gamma - 1) / (\gamma + 1))^{1/\gamma}.$$

According to prospect theory, this inflection point is the reference point. If the logistic curve was convex or concave on the entire interval from 0 to LE, then the reference point was set to LE or 0, respectively. Thus for each respondent the individual reference point was estimated, expressed in life years (absolute reference point) and relative to the remaining life expectancy (relative reference point). We assessed the frequency of the reference point codes, focus of attention codes and goal codes per life-year CE gamble. For those respondents who mentioned a goal with an explicitly stated associated period, the correlation between that period and the (relative) reference point was assessed. The correlation between coded goals (number of years associated with), age and the (absolute or relative) reference point (as dependent variable) was investigated with a linear regression procedure.

Results

The respondents were twenty-six females aged 18 to 72 years (mean age = 27, s.d. = 12) and nineteen males aged 19 to 61 years (mean age = 34, s.d. = 4). They were educated to at least high school standard. About 50% of the respondents were university students, and 25% of the respondents had children under the age of eighteen years.

Proportional match

The highest mean PM (0.54) was observed for life-year gambles involving short periods of survival. The lowest mean PM (.29) was observed for life-year gambles involving long periods of survival (see Figure 3.2). In other words, respondents on average behaved in a risk-seeking way with respect to life-year gambles involving short periods of survival, and in a risk-averse way for all other gambles. Risk-averse behavior was strongest for life-year gambles involving long periods of survival. Figure 3.2 shows the mean PM per elicited CE.

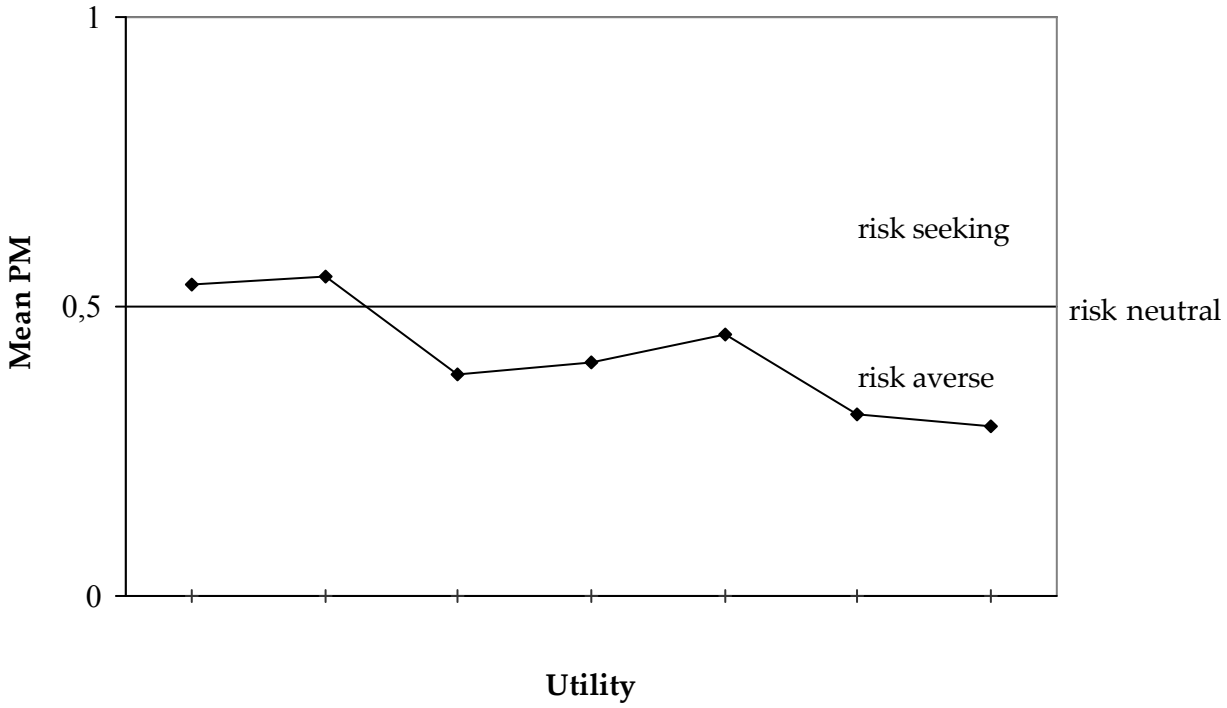

FIGURE 3.2. Mean proportional match (PM) per CE, PM > 0.5 denotes risk-seeking behavior, PM < 0.5 denotes risk-averse behavior, PM = 0.5 signifies risk-neutral behavior.

Logistic curve

The logistic curve provided a good fit for most patients, with an average explained variance of 0.927 (range 0.625 to 0.999). The estimated average absolute reference point was 14.3 years (s.d. = 19) and the average relative reference point was .29 (s.d. = .36).

Qualitative data

The analysis reported here is based on qualitative data from the CE. Gender, age and the CE that was valued, identify the quotes, which we use to illustrate our findings. It was difficult for some respondents to combine verbalization with the task, this was apparent in the reticence of several respondents to verbalize during the task.

For life-year gambles involving short periods of survival, several respondents indicated that they found all three outcomes unattractive. Regarding life-year gambles involving long periods of survival, some respondents viewed the low outcome of the gamble to be a satisfactory survival period. Some respondents even stated that they did not want to live as long as their life expectancy. In advanced age, they anticipated, for example, disease, handicaps, or 'problems of old age' ("The years between 50 and 80 aren't that much fun anyway, as your health will rapidly decline", female, 31, CE87.5). Frequently, choices were based on a feeling ("I find it hard to explain, it is a feeling", female, 39, CE25).

Reference point

With respect to all CE's, the offered CE was the outcome in the gamble that most frequently served as reference point (61%), followed by the low outcome (22%) and the high outcome (17%) (See Figure 3.3). During the CE50, few codings of "reference point" were deducible.

From Figure 3.3, it appears that a shift occurred in the outcome that indicated the reference point within the gamble (i.e. high and low outcomes). For the life-year gambles involving short periods of survival (e.g. CE12.5 and CE25), the high outcome of the gamble frequently served as the reference point. For the life-year gambles involving long periods of survival (e.g. CE75 and CE87.5), the low outcome of the gamble frequently served as the reference point (see Figure 3.3). Thus, the perceived reference point appeared (partly) dependent on the time horizon. Furthermore, 'death within a week' (the low outcome in the CE12.5, CE25, and CE50) never served as a reference point.

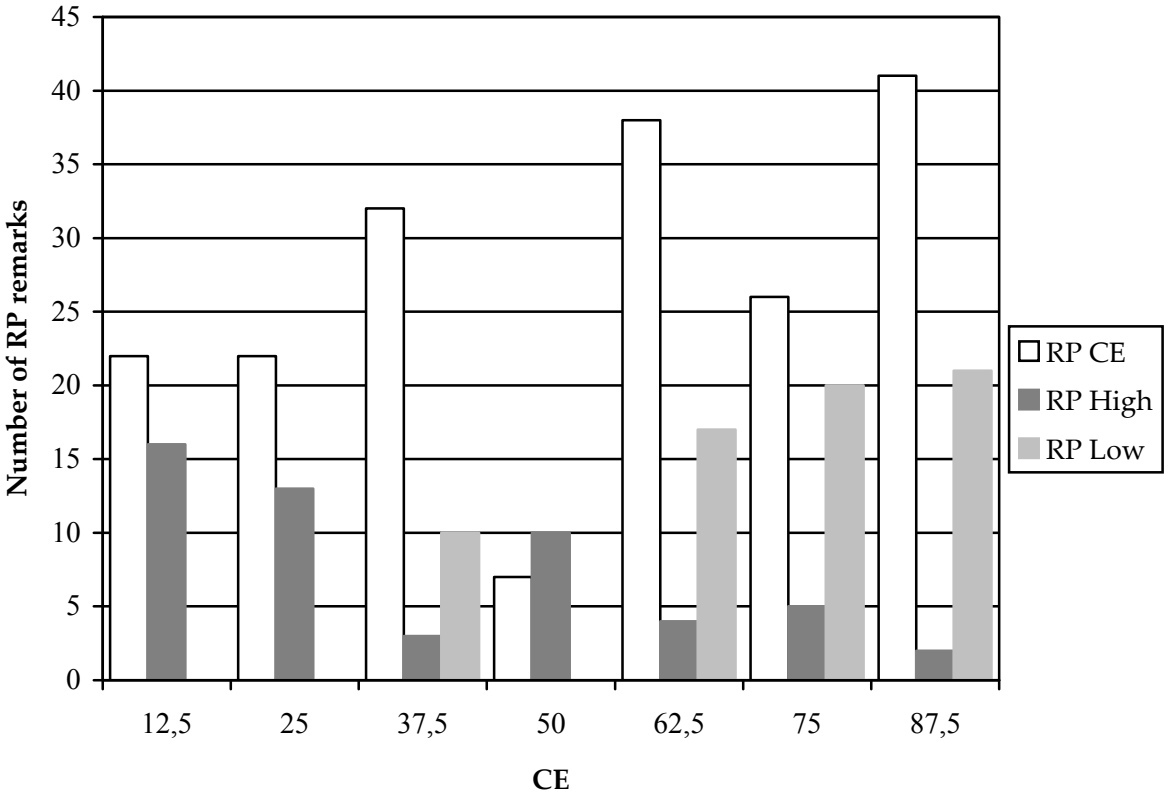

FIGURE 3.3. Frequency of deduced reference points (RP) shown per CE. Within the gamble, a shift was observed between reference points that appeared dependent on the time horizon (i.e. future life years involved in the life-year gamble).

Focus of attention

Respondents mentioned the offered CE most often (82%), followed by the high outcome (10%) and the low outcome (8%). Additionally, the offered CE was compared to most often. In 26% of all comparisons made (two-outcome comparisons and three-outcome comparisons), the offered CE was compared to the high outcome. Additionally, in 26% of all comparisons made, the offered CE was compared to the low outcome. Comparisons between the high and low outcome were made infrequently (8%). Comparing all three outcomes was done most frequently (40%).

A shift occurred in the frequency of comparing the offered CE to the high outcome as opposed to the frequency of comparing the offered CE to the low outcome (see Figure 3.4). As was the case for the reference point, the focus of attention seemed to be (partly) dependent on the time horizon. Regarding life-year gambles involving short periods of survival, the focus *within the gamble* lay mostly on the high outcome. This outcome was compared to the offered CE more frequently. Regarding life-year gambles involving long periods of survival, the focus *within the gamble* lay mostly on the low outcome. This outcome was compared to the offered CE more frequently.

Goals

Most respondents mentioned more than one goal during the interview. Goals were grouped into five categories: 1) Unspecified goals ("*I think if I live for that many years, I will still be relatively young, but can accomplish all I want to.*", female, 19, CE50), 2) Career-related goals ("*If I live for only three more years, I won't even be able to get my PhD.*", female, 31, CE50), 3) Retirement-related goals ("*I have to work until I am 65 and I want to enjoy more than one year of my retirement.*", male, 39, CE50), 4) Child-related goals ("*I won't consider my life a success if I cannot start a family; I could not do that if I only live until I am 39.*", male, 23, CE25), or 5) Miscellaneous other ("*If I choose this one, then I will be*

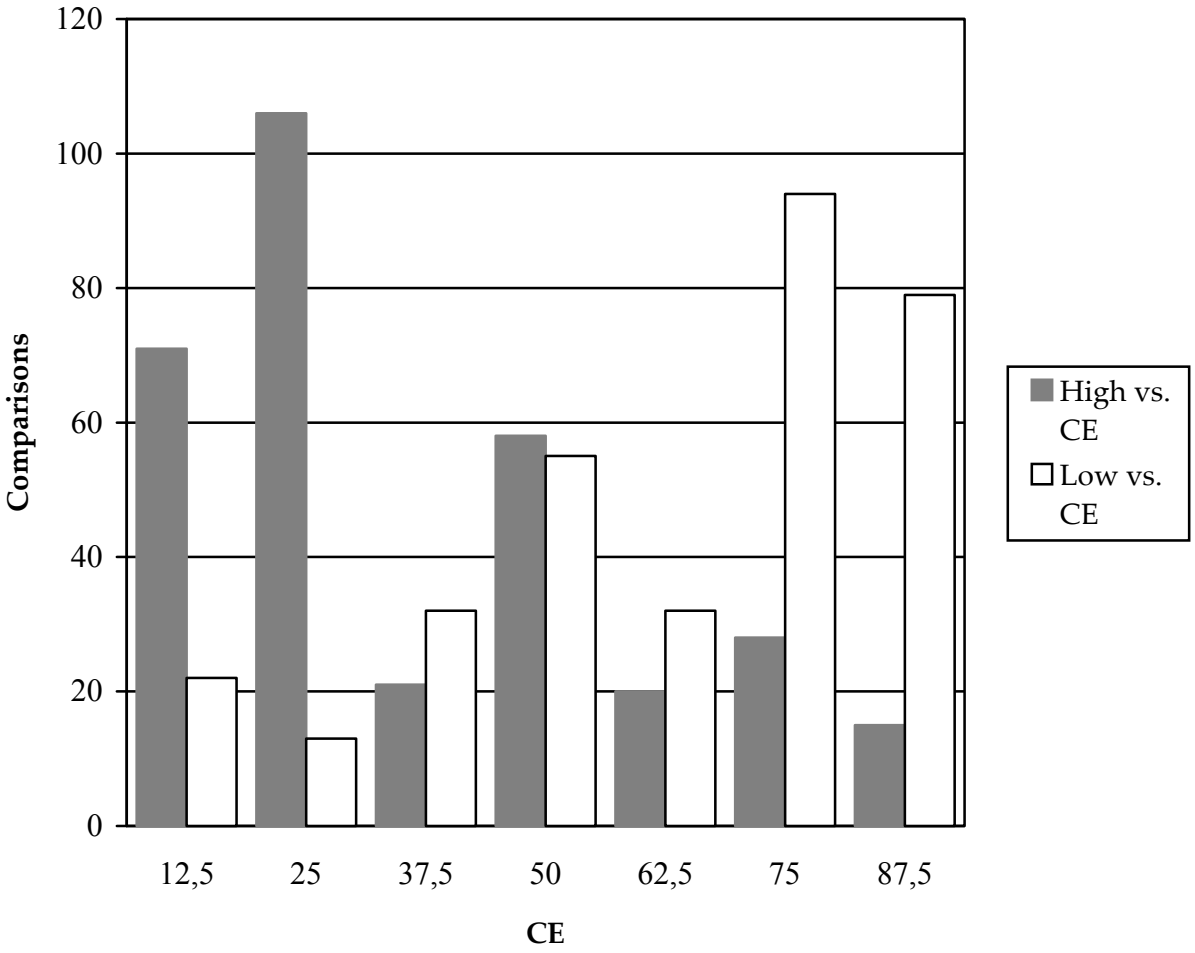

FIGURE 3.4. The frequency of two-outcome comparisons made with the offered CE, representing the focus of attention within the gamble. Again a shift was observed.

able to take care of my dog until she dies. If my dog dies, I won't care so much about living anymore." female, 72, CE50).

Respondents mostly used the period of time stated in the offered CE to assess if realization of a goal was possible (81%). The high outcome (12%) and low outcome (7%) were infrequently used to assess the possibility or impossibility of goal realization. However, if these outcomes were used to assess goal realization, again it was in a similar pattern depending on the time horizon. Mostly for the CE50 and CE75, the *possibility* of realizing a goal with respect to the offered CE, and to a lesser

extent with respect to the high outcome, was mentioned, but this was not the case with respect to the low outcome. On the other hand, mostly during the CE25 and CE50, the *impossibility* of realizing a goal was mentioned with respect to the offered CE and low outcome, but not with respect to the high outcome. Thus, again a shift occurred, depending on whether the life-year CE gamble involved trade-offs with long or short durations of survival. Goals were more frequently mentioned for nearby years than for the more remote life years.

Seventeen respondents mentioned a goal with an explicitly stated associated period, like "I need twenty years to raise my children." A linear regression showed that the explicitly stated periods are strongly related to the absolute ($p < .001$) and relative reference point ($p < .001$) that were indirectly obtained from the quantitative analysis, see Figure 3.5. The explained variance of the absolute reference point by the goal-related period was $R^2 = .66$. The seven respondents with a relative reference point of 0 (death within a week), had an average goal-related period of only 4 years. Conversely, the five respondents with a relative reference point of 1 had an average goal-related period of 33 years. Age showed a non-significant relation to the absolute reference point ($p = .19$) and relative reference point ($p = .82$).

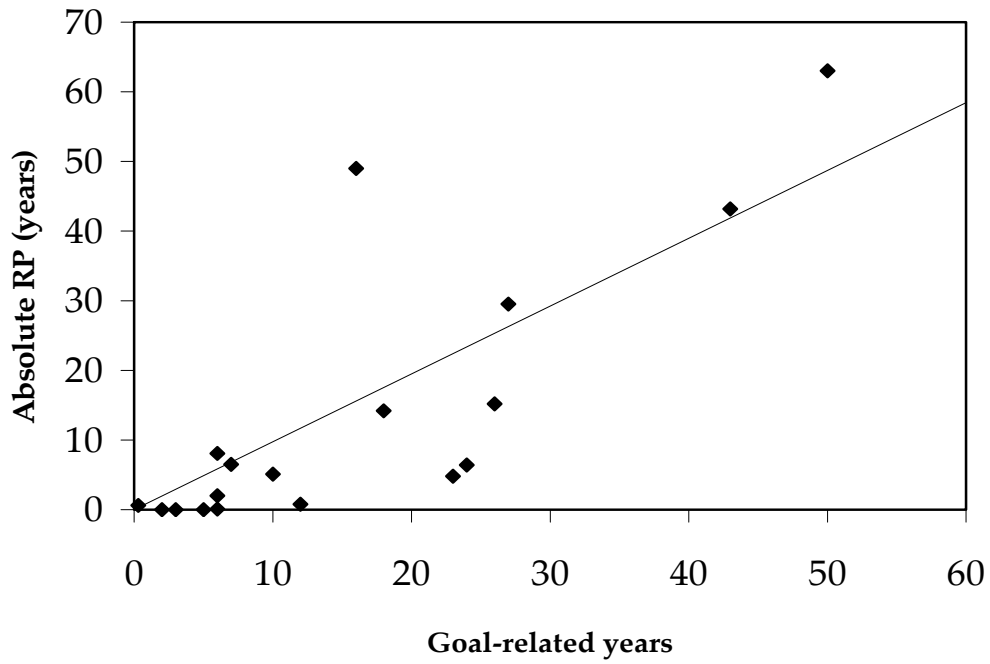

FIGURE 3.5. Scatter plot of the relation between goal-related years and the absolute reference point (RP) indirectly obtained from the quantitative analysis.

Discussion

The objective of this study was to assess the outcome in the CE gamble that seems closest to, or seems to include, the reference point. Additionally, the psychology behind the reference point was explored. To the best of our knowledge, no other studies have collected, in a systematic way, qualitative data on the process of valuing SG data that enabled study of the reference point. The data presented here provide evidence that the offered CE most frequently served as a reference point, and, thus, the life-year CE is most likely perceived as a mixed gamble in which the low outcome is perceived as a loss and the high outcome is perceived as a gain. Also an interaction was observed between the outcome that served as reference point and the time

horizon. A similar interaction was observed between the outcome that loomed largest and the time horizon. Frequently, some, or all outcomes were explicitly perceived as losses. Our qualitative findings argue that the CE life-year gamble is very likely not perceived as an all gains gamble, as has been suggested by Bleichrodt et al. (4). Quantitative data corroborate this finding, arguing that, contrary to a suggestion made by Bleichrodt et al., respondents behaved in a risk-seeking way for life-year CE gambles involving short periods of survival. They behaved in a risk-averse way for all other gambles, particularly so for gambles involving long durations of survival. Since prospect theory predicts risk-seeking behavior for losses, this provides indirect evidence that the reference point will not be zero life years. Indeed, in our quantitative data, only seven respondents showed a reference point of zero life years. Moreover, in our qualitative data, the outcome 'death within a week' never served as a reference point.

We propose two factors that may explain why the offered CE mostly served as a reference point: framing and goals. For all CEs, our qualitative data show that most attention was paid to the offered CE, which was mentioned and compared to most often. In the choice-bracketing search procedure, its value changes with every answer and, as a result, draws most attention. If the task is to be done properly, this is a logical consequence. That most attention was paid to the offered CE is a probable cause explaining why respondents mostly compared goals to that outcome, and to a lesser extent to the high and low outcomes of the gamble. As the offered CE attracts more attention, it renders itself nicely as a starting point for comparisons. Consequently, it serves well as a reference point due to framing of the question. We did not explicitly request information about the reference point and are, therefore, not capable to determine its exact location. Through the use of the choice-bracketing procedure we may have induced a changing reference point in the way one introduces a change in

the reference point by offering respondents a money amount to start with in money gambles.

However, framing cannot explain the shift in the outcomes that served as reference point, which occurred when the gambles involved trade-offs with relatively long and short durations of survival. We propose that goals caused the shift in outcomes that served as reference point because the outcomes closest to a goal-realizing period most likely served as the reference point. Respondents could experience some control over the offered CE, and, thus, the outcome was frequently closest to the goal-realizing period. Consequently, it was most often closest to the reference point. To live for the period in which goal realization is possible appears to underlie the motivational processes (29;43). This notion is supported by combining the quantitative and qualitative data that showed a strong relationship between the explicitly stated goal related period and the reference point indirectly deduced from the quantitative analysis. In non-health settings, it has been shown that the goal itself may serve as a reference point (44). Other research on goals presents related findings, such as Verhoef et al. They also observed a trend towards a lower absolute reference point with increasing age (29). We found no significant relation between age and the reference point. We argue that goals alter the perception of outcomes, as described by prospect theory, by influencing the reference point. This relationship is more apparent for the near future as opposed to the remote future, as goals are mostly set for the near future. Our findings support this. The preference curve shows that the reference point on average lies at 29% of the personal life expectancy and not the long term. Setting goals in the relatively near future enhances self-efficacy, and therefore, expectations of future success; as opposed to setting goals for life years in the long term, for which task mastery is low (44).

To summarize, if the life-year CE gamble involved short periods of time, life years were involved, for which goals were formulated. Consequently, attention shifted to the outcome involving the longest length of time, enhancing the likelihood of realizing set goals. As a result, risk taking increased for short-term gambles. For some respondents, the high outcome even served as the reference point. If a life-year CE gamble involved long durations of survival, few goals appeared set. Consequently, the prospect of goal realization was not in danger. As a result, the outcome closest to the goal-realizing period loomed largest, in this case the low outcome. For some respondents, the low outcome even served as the reference point. Mostly, people minimize risk taking if it is not required, such as for life-year CE gambles involving long durations of survival. Consequently, respondents take fewer risks. Some subjects did not even want to live as long as their remaining life expectancy, e.g. because they anticipated disease at advanced age. This could serve as a confounder in utility assessment.

Our findings also follow Lopes' SP/A (security-potential/aspiration) theory, a descriptive two-factor theory. The security-potential factor reflects the way that a person usually looks at risk, allowing for a perceptual change with respect to risk. This change depends on the second factor, the aspiration level. The aspiration level is a kind of reference point, but situational. It can be task dependent. Choices involve a balance between one's disposition towards risk (Lopes: degree of preference for Security versus Potential) and the opportunities of the situation, i.e. goals that can be realized during a number of years offered in an outcome (Lopes: Aspiration level) (45;46). The idea that goals strongly influence motivations has also been reported in other studies in other disciplines (39;43;47).

One of the limitations of our study was the order of the elicitations. The CE50 was always performed first. Subsequent CEs rely on the results (and possible errors) from the CE50 assessment, and this may lead to an upward or downward bias. To minimize the undesirable effect of error propagation, respondents were given two examples as well as an extensive verbal and written explanation beforehand. Due to verbalization and the presence of a researcher during the task, errors with respect to the task were noticeable and immediately corrected during the examples. We believe that due to these measures errors were minimized. However, a learning process was still observed. There were few "reference points" coded during the CE50, and much attention was paid to the offered CE. Moreover in the CE50, the high outcome (i.e. life expectancy) most frequently served as the reference point. The life expectancy appeared to influence the reference point when respondents first started the interview. A further limitation to our study was that about half of our sample consisted of university students. This may have implications for the extent to which our findings can be generalized. A further important point is that the findings are applicable only to the choice-bracketing elicitation method. If utilities had been derived using the matching method, these findings may have been different. Additionally, the think-aloud protocol is not without limitations. We were unable to code that which respondents did not verbalize.

Conclusions

On the basis of direct and indirect evidence, we conclude that the offered CE most often serves as the reference point and, therefore, that a life-year CE gamble is most likely perceived as a mixed gamble. This can have important consequences for economic evaluations, since for mixed gambles the corrected valuations differ from the uncorrected valuations (11;34). Framing and goals influence the outcome that served as reference point. Goals also cause an interaction between the time horizon and the attention paid to an outcome. We argue that motivational constructs should be incorporated into research of the reference point. Additionally, when patient preferences are involved in treatment decisions, it is desirable to address a patient's goals, because these influence preferences.

---

[i] Strictly speaking, to circumvent a Catch-22 situation, our analysis is based on inconsistent assumptions. First, assuming expected utility (EU), the utility of the certain outcome in each CE is calculated as the (untransformed) EU of the corresponding 50-50 gamble. Next, the inflection point of the thus estimated S-shaped utility curves is interpreted as the reference point assuming prospect theory. Considering the results of the linear regression analysis, however, we have confidence that this procedure results in a good estimate of the reference point.

Appendix 3A. Stimulus screen

Imagine you have a choice between 1, 2, and 3.

Choice 1

A gamble between:

-a 50% chance of living in good health for

.. years and .. month(s), and

-a 50% chance of dying within one week.

Choice 2

A guarantee (100%) that you will live in good health for

.. years and .. month(s).

Choice 3

I have no preference, for me choices 1 and 2 are equal.

Press, according to your choice 1, 2, or 3.

4

The construction of standard gamble utilities

The construction of standard gamble utilities.

S.M.C. van Osch, A.M. Stiggelbout

Health Economics: 2007; in press.

Abstract

Health effects for cost-effectiveness analysis are best measured in life years, with quality of life in each life year expressed in terms of utilities. The standard gamble (SG) has been the gold standard for utility measurement. However, the biases of probability weighting, loss aversion and scale compatibility have an inconclusive effect on SG utilities. We determined their effect on SG utilities using qualitative data to assess the reference point and the focus of attention. While thinking aloud, 45 healthy respondents provided SG utilities for six rheumatoid arthritis health states. Reference points, goals, and focuses of attention were coded. To assess the effect of scale compatibility, correlations were assessed between focus of attention and mean utility. The certain outcome served most frequently as reference point, and the SG was perceived as a mixed gamble. Goals were mostly related to this outcome. Scale compatibility led to a significant upward bias in utilities; attention was relatively more with the low outcome and this was positively correlated with mean utility. SG utilities should be corrected for loss aversion and probability weighting with the mixed correction formula proposed by prospect theory. Scale compatibility will still bias SG utilities, calling for research on ways to correct for this bias.

Introduction

In cost-effectiveness analyses, the costs associated with a health care intervention are compared to its benefits (health effects). It is commonly acknowledged that health effects are best measured in terms life years and that quality of life in each life year is best expressed in terms of utilities, that are then used as a weighting factor, yielding quality-adjusted life years (QALYs) (8;9;48). Based on normative expected-utility arguments, the standard gamble (SG) method has often been considered the gold standard for utility measurement because, unlike other elicitation methods, it incorporates risk. The standard gamble generally requires a respondent to compare the certainty of being in the heath state to be valued for the remaining life expectancy, with a gamble that offers a chance of optimal health for the remaining life expectancy but also entails a risk of immediate death. In the generally used probability equivalent of the SG, the respondent is asked to indicate at what probabilities of the gamble he or she would be indifferent to the choice between the health state and the gamble.

There is much empirical evidence demonstrating that expected utility is not descriptively valid, the three main reasons for this being probability weighting, loss aversion, and scale compatibility (6;7;11;13). These biases generally lead to SG utilities that are too high. The use of biased utilities may lead to biased resource allocation decisions and, therefore the joint effect of these biases should be minimized in the elicitation of utilities. This leaves a great need for more knowledge about these biases in health utility measurement, so as to adequately correct for them or to be aware of their effect on SG utilities (and health care choices) (4).

We will first discuss the biases of loss aversion and probability weighting, because they interact. Loss aversion implies that the disutility that a person experiences from

losses is significantly greater than the utility the person experiences from gains of the same absolute size (7). Losses weigh more heavily in decisions than gains do. Probability weighting entails that people weight probabilities in a nonlinear manner. Probability weighting will usually lead to upward bias for SG utilities (6). Cumulative prospect theory (PT) proposes to transform probabilities by using a rank-dependent probability-weighting function (5;49).

The weighting function that corrects for probability weighting and loss aversion depends on the perception of the gamble. The SG can be perceived as yielding all losses, all gains, or as being mixed (yielding both gains and losses), depending on the perceived reference point. SG utilities are only distorted by loss aversion if the SG is perceived as a mixed gamble. To adequately correct for probability weighting, or to know whether loss aversion influences SG utilities, knowledge about the location of the reference point is needed. However, PT incorporates no theory with respect to the reference point, and direct evidence for the location of the reference point is absent.

Hershey and Shoemaker, as well as Bleichrodt et al., reasoned that the certain outcome is fixed in SG elicitation and, therefore, may provide a salient reference point (6;7). Robinson et al. argued that a task instruction could lead to adopting a certain reference point (50). For example, if the instruction reads that respondents should imagine themselves 'to be' in the health state to be valued, i.e. the certain outcome, after which a medical intervention is to be followed that can result in success (a return to optimal health) or failure (death), this will probably cause respondents to adopt the certain outcome as the reference point. Robinson at el. indeed observed this instruction effect on the reference point in their study. The scale-compatibility bias results from the principle that the attributes of decision alternatives that are compatible with the response scale are weighted more heavily in decisions (6;23).

In the probability equivalent version of the SG, the response scale is a probability, and, thus, the individual will focus on the probability. The associated outcome will receive more weight. The problem is, as Bleichrodt argued, that the bias may hold equally well for the certain outcome probability, the good-outcome probability as for the bad-outcome probability (6). Scale compatibility leads to higher utilities if a respondent focuses on the outcome that involves the bad-outcome probability. The gamble would become less appealing compared to the certain outcome. To achieve indifference, the gamble must be made more appealing which is accomplished by increasing the good-outcome probability. Following a similar reasoning, scale compatibility would lead to lower utilities if a respondent focuses on the good-outcome probability, as then the gamble is found to be more appealing than the certain outcome. This bias leads to a preference for the gamble over the certain outcome. Indifference is achieved by decreasing the good-outcome probability. Thus, the scale-compatibility bias could lead to either higher or lower SG utilities. It is unknown whether one of the outcomes (probabilities) draws more attention than the other. Given that the probability of the certain outcome is not varied (p= 1), we assume that either an upward bias on SG utilities (due to focus on the bad outcome) or a downward bias on SG utilities (due to focus on the good outcome) will occur.

The purpose of this study was to further explore SG biases using qualitative data (using a 'think-aloud' protocol). The aims were twofold. The first aim was to locate the SG outcome that is the reference point or seems to lie closest to the reference point. In the rest of the paper, we speak of 'the outcome that serves as reference point'. With knowledge of the reference point, the proper correction method can be applied to counteract probability weighting and loss aversion (the latter if necessary). In an earlier study, we observed that the desire to attain goals (e.g. raise children, make a career) probably influences the perception of the reference point in certainty

equivalent standard gambles (51). In a non-health setting, one study even argued that goals served as the reference point (44). Therefore, we focused on goals in relation to the outcomes as well.

The second objective was to obtain an indication of whether scale compatibility results in a systematic bias upwards or downwards, i.e. whether respondents focus more on the bad outcome or the good outcome? To assess this point, we identified the focus of attention. Additionally, we aimed to verify that a main focus on a bad outcome or good outcome leads to higher or lower utilities respectively. Furthermore, relevant themes raised by respondents during the experiment, that could result in a biased utility, were also taken into account.

Methods

Sample

Forty-five respondents, recruited via newspaper advertisements and pamphlets, participated in the study. Each respondent was paid € 22.50 for participation in two interviews (conducted by the first author, SMCvO); one of which is the topic of this paper. The interview took 45 minutes on average to complete. No specific sample criteria were applied.

Procedure

Six rheumatoid arthritis health state descriptions were selected from those given by rheumatic patients in the Rheumatoid Arthritis Patients In Training study (52). Descriptions were according to the EQ-5D system, a multi-attribute health-utility system. The EQ-5D system includes five dimensions of health (mobility, self-care,

usual activities, pain/discomfort, and anxiety/depression). Each dimension comprises three levels (no problems, some/moderate problems, and extreme problems). A unique EQ-5D health state combines one level from each of the five dimensions. The health states were chosen to cover the utility continuum (0 to 1), using corresponding EQ-5D valuations from the general public based on the TTO (25). We used the EQ-5D health state descriptions; 21232 (I, EQ-5D index of .09), 22322 (II, EQ-5D index of .19), 21321 (III, EQ-5D index of .36), 21222 (IV, EQ-5D index of .62), 21211 (V, EQ-5D index of .81), and 21111 (VI, EQ-5D index of .85).

The SG was written using the program Ci3 (15). All elicitations were based on the choice-bracketing search procedure, i.e. a series of ping-pong questions. In total, six SGs were elicited, one for each rheumatoid arthritis health state. The order of elicitations was randomized. The experiment started with a written explanation of RA. An oral and written explanation of the SG followed, after which two examples (practice tasks) were given. We took great care regarding the instruction in order to avoid an instruction effect on the reference point as mentioned in the introduction (see Appendix 4A). Two 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 optimal 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 life expectancy as derived from Dutch life tables (16). At any time during an elicitation, it was possible for respondents to take a break or check earlier answers within that elicitation and change these. Elicitations ended when respondents indicated that they valued two options equally. All respondents were instructed to think aloud during the interviews without paying attention to pronouncing grammatically correct sentences. They received two examples to practice verbalizing before each task. If a subject became silent during the

task additional instructions were given to think aloud. Thinking aloud has been reported to only affect the time needed to complete the task (40), and not the task itself. Interviews were taped and transcribed.

Coding

Coding involved initial familiarization with the qualitative data, sorting and indexing of relevant themes. These steps were iterative and not strictly consecutive. Two independent coders each coded the first half of the reports to resolve any ambiguities. Observed differences in coding were discussed and for these a consensus coding was reached. One coder coded the remaining interviews for themes deduced from this process.

A "reference point" was coded if a comparison relative to a point of view was formulated, i.e. if respondents used one of the three outcomes of the SG as a starting point to indicate the difference (gain or loss) with one outcome, or with both other outcomes. The three outcomes that could serve as the reference point were: the good outcome of the gamble, the bad outcome of the gamble, and the certain outcome of the gamble. The latter was the RA health state that was valued. We coded the bad outcome as the "reference point" if it was taken as starting point, and one of both other outcomes was/were labeled as gains relative to this bad outcome. It was also coded as "reference point" if respondents indicated that they perceived this outcome, i.e. death within a week, as the current reference situation from which the SG was perceived. For example, "I have nothing and can gain 40 years of optimal health or have rheumatism if I choose option 2". Similarly, we coded the certain outcome as "reference point" if the outcome was taken as the starting point, e.g., "Either I continue living with these complaints, or I will have an operation and the operation will succeed, or it will not and I will die." We coded the good outcome as the "reference

point" if that outcome was taken as the starting point, e.g., "I will stay healthy, or lose it all, or end up with rheumatism."

A "goal" was coded if a statement was made regarding the realization of a goal with respect to an outcome. In other words, if the respondent referred to an outcome and assessed that outcome to be sufficient or insufficient for the realization of the goal.

A "focus of attention" was coded when a respondent mentioned the good or bad outcome including its probability ("There is a 50% chance to die within a week."). Additionally, "focus of attention" was coded when the good or bad outcome, including its probability, was compared to another outcome ("There is a 50% chance to die within a week and a 50% chance to live for 40 years."). There is no point of view identifiable in this comparison. We assumed that the more frequently an outcome was mentioned or compared, the more attention a respondent paid to that outcome.

Data analysis

Values of 0 and 1 were assigned to 'death within a week' and 'optimal health' respectively, and the probability of the good outcome at the point of indifference was then taken as the utility value of the RA health state. The frequency of the coded reference points was assessed across respondents. The frequency of the coded goals was assessed across respondents. We determined the relative focus of attention for each respondent by calculating whether the focus was relatively more with the good or with the bad outcome of the gamble (Total focus on bad outcome / (Total focus on bad outcome + Total focus on good outcome)). The effect of scale compatibility on mean SG utilities was tested by calculating Pearson's *R* correlations between the relative focus of attention and the total mean utility for the six health states.

Results

The respondents consisted of 26 women aged 18 to 72 years (mean age = 27, s.d. = 12) and 19 men aged 19 to 61 years (mean age = 34, s.d. = 14). All respondents had received at least a high-school level of education. About 50% of the respondents were university students (mostly in biomedical science or medicine), and 25% of the respondents had children. The aggregate utility values for health states were; mean utility for health state I = .54 (s.d. = .28), mean II = .57 (s.d. = .29), mean III = .69 (s.d. = .26), mean IV = .72 (s.d. = .24), mean V = .84, (s.d. = .21), mean VI = .87 (s.d. = .17). Gender, age and the health state that is valued identify the quotes which we use to illustrate our findings. It was difficult for some respondents to combine verbalization with the task, as was apparent from the reticence of several respondents to verbalize during the task.

Reference point

The certain outcome (RA) most often served as reference point (52% of reference point codings), e.g.

> *"It is quite troublesome that I am not able to perform my daily activities .. I would choose for the operation so to say … I know I will be handicapped for life if I cannot perform my daily activities. And I will have a 10% chance that I will be able to do that."* (male, 61, II).

This was closely followed by the high outcome (LE) (45.5% of reference point codings), e.g.

"40% Chance to live. I do not like to gamble with my life, but I must be a good life. Yes, I think it would be equally bad to have a 60% chance to die within a week or to lose my independence" (female, 22, II).

The bad outcome (death within a week) hardly ever served as reference point (2.5 % of reference point codings), e.g.:

"A 50% to die within a week. … Well, I die tomorrow, then I would rather live somewhat longer, but then with pain." (female, 20, IV).

Goals

Goals were mostly mentioned with respect to the certain outcome (77%), and to a lesser extent with respect to the bad outcome (14%) or the good outcome (9%). Most frequently, respondents indicated not being able to achieve their goal(s) (86%), usually with respect to the more severe health states. Most respondents mentioned more than one goal during the interview. A concern about whether goal aspiration was still possible was often expressed as a desire: 1) to either live life in a certain way (preferably in the same way as now), and/or 2) to achieve specific goals that had been set. Regarding the first concern, respondents considered the impact of the RA health state on their way of life, e.g. the ability to take part in sports, to walk, to be independent.

"What's the use of living that long without being able to do my thing?" (male, 41, III)

"A 10% chance of dying. I would certainly take the risk. To be released from the pain and the fact that I can again just do everything. That I will be able to walk." (female, 23, I)

With respect to the second concern, this was mostly related to the gamble, specifically, to the bad outcome of the gamble.

> *"I choose option 1. For I want the security to live, because I still have to care for my children … My youngest is 10, and then I have a daughter who is 12, and a son who is 14 and a daughter who is 20 … I am very involved in my children's lives."* (female, 41, II)

> *"Yes, well, I constantly think that you are in a sort of family situation. And well, basically, it is the same argumentation over again ... Yes, well wife and children and everything … just that you can't leave that situation and you would rather put up with the pain than take the chance of dying."* (male, 23, I)

Focus of attention

Respondents barely mentioned the probability (100%) of the certain outcome. They mainly focused on the bad outcome of the gamble. Figure 4.1 presents the focus of attention per health state for the good and bad outcomes. It shows that for all health states the bad outcome drew relatively more attention than the good outcome. Additionally, focus of attention showed a significant correlation with the mean SG utility. The more the focus was with the bad outcome of the gamble the higher the mean utility was (Pearson's R correlation = .40, $p < .05$). The more the focus was with the good outcome the lower the mean utility was (Pearson's R correlation = -.40, $p < .05$).

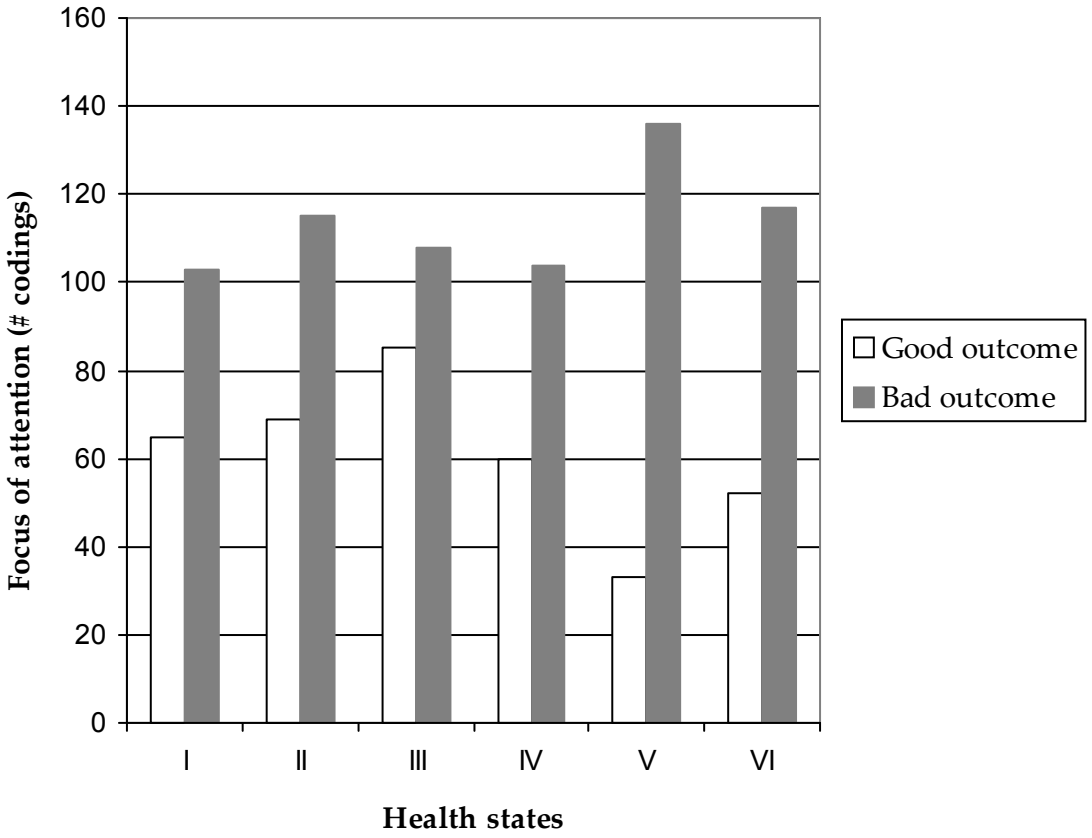

FIGURE 4.1. The frequency of focus of attention codings is depicted for the good outcome and the bad outcome of the gamble per health state. The bad outcome drew most attention.

Other themes

Other themes that may have had an effect on utilities were anticipated adaptation, and maximal endurable time (MET). Anticipated adaptation describes how respondents expected to adapt to the RA health state in due time, with a positive impact on utilities. This theme was considered fairly frequently. We distinguished between psychological and physical adaptation (53). The first is the ability to emotionally adapt to a health state, e.g. a shift in personal interests.

> *"I think, it may be a question of wanting … you may think in the beginning that you cannot do it, so I, uh … I wouldn't be up to it. To a certain point in time when you will accept it … being so dependent, you can get used to that, I think."* (female, 37, II)

Physical adaptation concerns the physical aspect of adaptation, e.g. taking painkillers.

> *"I don't have much knowledge about pain, but isn't there medication for that? So basically, you can live with the pain."* (female, 20, I)

MET (maximal endurable time) is the concept that when people live in a health state involving a low quality of life, at first they evaluate it positively, but after a certain period (MET), any additional time spent in that health state is viewed negatively. Consequently, death is preferable after that period (54;55). METs were considered sometimes.

> *"I wish I could live for ten years with these problems and that it would be over after that … to die within a week is far from attractive, but neither is going on like this for another 55 years."* (female, 26, II)

Discussion

If a person behaves perfectly in agreement with expected utility theory, then the SG method yields unbiased utility values. Empirical evidence of inconsistencies in SG valuations is well documented and has led to so-called non-expected utility theories, e.g. PT, in which biases such as probability weighting and loss aversion are described, and from which corrections can be deduced. Correcting for biases or knowledge about the effect of biases will increase the descriptive validity of the SG without sacrificing its normative validity (56). The application of qualitative techniques in our study yielded rich data on SG utilities and provided new insights into relevant biases.

Our qualitative data showed that the certain outcome most often served as the reference point. There are theoretical arguments, based on prospect theory, favoring the mixed-correction of the SG for which the certain outcome served as reference point. Another study reported the highest convergent validity between the TTO and SG if the certain outcome was used as the reference point (34). Baker and Robinson found evidence of the certain outcome serving as the reference point as well, although they did not recognize it as such. In their study, respondents framed the SG as being in the intermediate health state and choosing for the operation with a success-probability and failure-probability (57). We specifically aimed to avoid instructing respondents to adopt an outcome as the reference point. Many studies, although probably unintentionally, instruct their respondents to adopt the certain outcome as the reference point. If one wants to compare health effects between studies, as is common in cost-effectiveness analysis, it is preferable that SG utilities are elicited with the same outcome as the reference point, e.g. the health state to be valued. The biases probability weighting and loss aversion then have similar effects on SG utilities, and the proper correction method is known. An instruction may be used to this effect, in which a reference point is implied.

Not surprisingly, the low outcome of the gamble, i.e. death within a week, hardly ever served as reference point. This outcome is too far from the present health status and does not enjoy the benefits of the certain outcome. Although no theory supports the finding, the high outcome quite frequently served as the reference point. A logical explanation is that respondents involved their status quo, which likely was most equal to the high outcome, i.e. optimal health during the remaining life expectancy. After all, the situation in which the respondent has a 100% chance of having the high outcome was, most likely, their actual situation before they started the interview. This finding may be viewed as an anchoring effect.

Several goals were mentioned, mostly through concerns of not being able to achieve goals. Goals were mostly referred to with respect to the certain outcome. The relationship between goals and reference point is natural. Goals influence motivation, and thus, decision making. Lopes for example assigns a role to the aspiration level, e.g. goals, in risky decision making (58). Our findings support the hypothesis that the certain outcome most often served as the reference point.

The certain outcome probability, which does not change during the task, drew the least attention. The bad-outcome probability of the gamble appeared to draw most attention. Consequently, the theorized upwards effect of the scale-compatibility bias was observed. SG utilities were significantly higher if respondents focused more on the bad-outcome probability than on the good-outcome probability of the gamble. This provides new evidence as to why SG utilities are generally observed to be too high. It has been observed before in the medical field that people exhibit a systematic dislike for risk. Other studies report that risk avoiders generally focus on the worst outcome whereas risk seekers focus on the best outcome (59-61). Most respondents behave risk avoiding for the SG, as was the case in our study. And as expected, respondents mostly perceived the SG as a mixed gamble, and the low outcome of the gamble was therefore perceived as a loss, and over-all received most attention. Loss aversion states that a loss looms larger than a gain, which may be interpreted as support for the finding that the focus of attention lies relatively more on the low outcome.

Other themes that were relevant to our subjects were anticipated adaptation and MET. These themes may have an effect on other elicitation methods as well. The first, adaptation, may lead to higher utilities for respondents who expect to adapt to the health state. A study by Baker & Robinson reported that respondents considered

anticipated adaptation during the SG as well (57). Damschroder et al. have shown that asking respondents to perform a simple adaptation exercise, in which they were primed to anticipate their adaptation, indeed led to higher values in a person tradeoff method (62). Some respondents exhibited MET. This can lead to lower utilities, as the certain outcome is then found to be less attractive than the death outcome. This is especially so for more severely impaired health states for which maximal endurable time is likely to arise.

A limitation of our study was that approximately half of our sample consisted of university students. As a result, the findings may not generalize to the general public. A further important point is that the findings are applicable only to the choice-bracketing elicitation method. If utilities had been derived using another elicitation method (e.g. matching), these findings might have been different. Additionally, the think-aloud protocol is not without limitations. We were unable to code that which respondents did not verbalize.

Conclusions

We argue that the certain outcome, i.e. the health state to be valued, most often served as the reference point, and, therefore, a SG is most likely perceived as a mixed gamble. This can have important consequences for cost-effectiveness analyses, such as correcting preference-based utilities for loss aversion and probability weighting with the appropriate mixed correction formula (11;34). Additionally, we observed that respondents mostly focused on the low outcome of the gamble. Consequently, scale compatibility will have led to upward biases in the SG utilities.

Appendix 4A. Introduction and stimulus screen

Introduction screen

In a moment you will see an example with three possible choices. If you choose number 1, then you are choosing to definitely (100%) live for your total life expectancy with the severity of rheumatism which has been described. If you choose number 2, then you are taking a gamble: on the one hand you have the chance (…%) of living your total life expectancy in optimal health, but on the other hand, you have a chance (…%) of dying within a week. If you choose number 3, then you think that choices 1 and 2 have equal value. You don't have a preference for either choice.

Stimulus screen

Choice 1

Definitely (100%) live with rheumatoid arthritis for the rest of your life expectancy. You have:

- some problems with walking
- no problem washing and dressing yourself
- some problems with daily activities
- very severe pain or other symptoms
- you are fairly fearful or gloomy.

Choice 2

A gamble between:

- ..% chance of living in good health for the rest of your life expectancy

OR

- ..% chance of dying within a week.

Choice 3

I have no preference, for me choices 1 and 2 are equal.

5

The construction of time trade-off utilities

The construction of time trade-off utilities.

S.M.C. van Osch, A.M. Stiggelbout

Submitted for publication

Abstract

The time trade-off (TTO) is an important method to elicit health state utilities. As all methods to elicit utilities do, it suffers from biases. Additionally, providing a health state with a label (e.g. cancer) has been known to possibly influence utilities. The purpose of this study is to obtain a better understanding of how TTO utilities are constructed and to discuss how biases and labeling may influence patient and public values. The first experiment used qualitative data to explore biases and themes that influence TTO utilities. Forty-five respondents valued six rheumatoid arthritis (RA) health states using a TTO, while thinking aloud. We found evidence of the biases scale compatibility and loss aversion. Qualitative data showed that respondents considered themes such as anticipated adaptation, life goals, and maximal endurable time. Effects of these themes on TTO utilities are discussed. Labeling of the health states appeared to influence utilities. In a second experiment, therefore, 84 respondents divided into two groups valued six RA health states either labeled or not labeled. Labeled health states were valued higher than unlabeled ones were. Labeling may have the effect that healthy subjects, like patients, do not use the whole utility continuum. Labeling a health state may lower discrepancies between patient and public estimates of quality of life.

Introduction

The Time trade-off (TTO) is an important method to elicit health state utilities that are used in cost-utility analysis (10). The TTO assesses the period in good health that is equivalent to a period in bad health. Several studies have presented empirical evidence demonstrating that TTO utilities better reflect individual preferences for health than standard gamble (SG) utilities, due to biases in the latter method (6;34;63). Nevertheless, some biases are known to influence TTO utilities as well. The most well known biases that affect TTO utilities are scale compatibility, utility curvature, and loss aversion (6). These biases play a role in the differences in utilities found between patients and the general public. These differences are the topic of active debate and research (53;64;65), and we will argue that part of these differences may be explained by loss aversion.

Scale compatibility refers to the finding that the higher the compatibility of a characteristic is with the response scale that is used, the more attention and weight an individual will give to that characteristic (6;13). The TTO response scale is in years and, thus, the duration of the state is likely to receive more attention than the health status. Consequently, scale compatibility will lead to higher utilities as people are reluctant to give up life years. The bias of utility curvature results from the assumption underlying the TTO that the utility for life years is linear. For most people this assumption is incorrect, as nearby years are valued more than remote years. Because remote years are more easily traded in the TTO method, this leads to lower utilities (6). Loss aversion refers to the finding that people are more sensitive to losses than they are to gains (11). In other words, the location of the reference point, or point of view, in the TTO is relevant to determine which outcome is perceived as a loss. Bleichrodt argued that the impaired health state is used as the reference point. The

TTO asks how many life years a person is willing to give up in order to *regain* optimal health. Thus life years are viewed as losses, leading to higher TTO utilities (6). Individual differences have indeed been found due to the point of view of the individual that values a health state, and these will likely be related to loss aversion (53;66;67). The differences in utilities between the general public and patients have mostly been attributed to psychological processes that patients have experienced, e.g. adaptation to and/or coping with a health state. Another explanation may be a different reference point. If the reference point for patients is the health state to be valued, whereas for a healthy person it is the state of good health, the bias loss aversion will operate differently for the two groups.

Another factor that may influence utilities and that, strictly speaking, is not a bias, is labeling. A study by Gerard et al. showed a shift in valuation that was partly due to the label 'cancer' that was given to the description (65). They argued that the possible connotations of the word 'cancer' could have led to lower valuation than was the case for non-labeled health states. Patients have their implicit labels, whereas in studies in the general public unlabelled health states from classification systems are commonly used.

This study aims for a better understanding of how TTO utilities are constructed. It is based on the increased interest in cognitive aspects of quality of life assessment (57;68;69). This study consists of two experiments. In the first experiment, the biases that influence the TTO are studied qualitatively. Qualitative data provide better insight into the construction of TTO utilities and possible themes and biases that influence it. This in turn sheds new light on discrepancies between utilities of patients and the general public. The second experiment assesses the possible effect of labeling. It further compares the derived TTO utilities with valuations from the EQ-5D tariff.

The purpose of this experiment was to test the impact of the labeling of health states on TTO utilities by healthy subjects.

Experiment 1

Methods

Procedure

Forty-five respondents were recruited through newspaper ads and pamphlets. Respondents were paid € 22.50 for participation in two interviews, one of which is the topic of this paper. Six RA health state descriptions were selected from the descriptions given by rheumatic patients in the Rheumatoid Arthritis Patients In Training study (24). Descriptions were taken from the EQ-5D system. The health states were chosen so that they would cover the utility continuum (0 – 1), using corresponding EQ-5D valuations based on the TTO (25).

We used the EQ-5D health state descriptions; 21232 (I, UK EQ-5D index = .09 (Dutch tariff = .17)), 22322 (II, EQ-5D index =.19 (.31)), 21321 (III, EQ-5D index = .36 (.52)), 21222 (IV, EQ-5D index = .62 (.65)), 21211 (V, EQ-5D index = .81(.86)), and 21111 (VI, EQ-5D index = .85 (.89)). The TTO was written using the computer program Ci3 (26). The interview took 45 minutes on average to complete. A researcher was present during all elicitations.

The experiment started with a written explanation of RA. A verbal and written explanation of the TTO followed, after which two examples were given. At any time during an elicitation, it was possible for respondents to take a break or check earlier

answers within that elicitation and change these. Elicitations ended when respondents indicated that they valued two options equally. In total, six TTO's were performed, one elicitation for each RA health state. The order of elicitations was randomized. Respondents were offered the choice between either a RA health state during the remaining life expectancy (LE) and a healthy life for period x ($x \leq LE$). Period x (offered number of years in optimal health) was varied (ping-pong) until indifference resulted. LE was based on a respondent's remaining life expectancy derived from Dutch life tables (28). Additionally, familiarity of the respondent with RA was assessed by asking whether they knew someone in the inner circle of acquaintances that suffered from RA.

All respondents were instructed to think aloud during the TTO interviews. Thinking aloud, i.e. verbalization, has been reported to affect only the time needed to complete the task (40). All qualitative data was taped and transcribed.

Coding

Initial familiarization with the qualitative data took place, this was followed by sorting, and indexing of other themes. These steps were iterative and not strictly consecutive. We focused on biases reported in the literature, and other themes relevant for TTO utilities. Two independent coders each coded the first quarter of the reports to resolve any ambiguities. Observed differences in coding were discussed and for these a consensus coding was reached. One coder coded the remaining interviews for themes based on this consensus.

In order to qualitatively determine the effect of scale compatibility, we planned to assess whether the focus lay more on life years than on the quality of life of the health state that is valued. Given the (ping-pong) task in the TTO, however, it turned

out that respondents overwhelmingly focused on life years. We therefore decided not to code this further. To assess the effect of loss aversion we coded whether outcomes were labeled as gains or losses, and explored the data for reference points (34). “Reference point" was coded if a point of view was formulated, i.e. respondents used an outcome as a starting point to indicate or calculate the difference with another outcome and, thus, indicated the perception of the outcomes as a loss or a gain.

Possible outcomes that could serve as reference point were the RA health state for the remaining life expectancy, period x in optimal health, or current health state (remaining life expectancy in optimal health, i.e. status quo). The reference point RA for the remaining life expectancy was coded as reference point if respondents adopted RA as their starting point and spoke of either regaining optimal health and/or being deprived of future life years. The outcome “Living for period x in optimal health” was coded as reference point if respondents used this outcome as starting point and spoke of gaining future life years and/or giving up optimal health or wishing to stay in optimal health. “Current health state” was coded as reference point if it was used as starting point and other outcomes were viewed as losses relative to this outcome. The current health state (living for optimal health for the remaining life expectancy) was not an option in TTO.

Other themes that were coded were maximal endurable time (54), anticipated adaptation (62), explicitly not anticipating adaptation, and anticipating problems related to old age. Furthermore a "goal" was coded if a statement was made regarding the realization of a goal with respect to one or both outcomes (34).

Data Analysis

For the qualitative data, we assessed the number of remarks on a theme made by that respondent per health state that was valued. Correlations between coded themes and utilities were assessed for all health states. To give an example, assuming that loss aversion leads to higher utilities: We expected a positive correlation between the utility on the one hand, and the number of times that the RA health state was used as reference point. Using a regression analysis (enter method) with TTO utility as dependent variable, we determined themes that correlated significantly with utilities.

Results

The respondents consisted of 26 women aged 18 to 72 years (mean age = 27, s.d. = 12) and 19 men aged 19 to 61 years (mean age = 34, s.d. = 14). All respondents were educated with at least high-school level. About 50% of the respondents were university students, and 25% of the respondents had children.

The analysis reported here is based on qualitative data from the TTO. Quotes, which are used to illustrate our findings, are identified by gender, respondent age and the health state that is valued (I to VI). It was difficult for some respondents to combine verbalization and the tradeoff task, which was apparent from the reticence of some respondents to verbalize during the TTO. Respondents considered a wide array of information during the valuation task, e.g. the impact of RA on their life and/or a shorter remaining LE:

"What I can do in 52 years with rheumatism, I can do in 20 years without rheumatism." (male, 23, V)

Loss aversion

The reference point that was most often used was the RA health state that was to be valued. It was used as reference point in 201 times out of the total of 317 reference points that were deduced at group level (63%). To live in the optimal health state for x years was used as the reference point in 82 out of the 317 instances (26%). The current health state was used as reference point in 11% (34) of the cases. Use of the RA health state as reference point correlated positively with utilities ($r = .40$, $p < .01$, see Table 5.1). The use of the optimal health state for x years as reference point showed no relation with utilities ($r = .14$, $p = .35$).

Utility curvature and anticipated problems in old age

Some subjects considered potential health problems in old age in their decision. This occurred if the offered number of years in good health was close to the respondent's remaining life expectancy. Obviously, this was more frequently observed for the less severe health states. Of the 14 remarks that were made on that theme, 93% occurred during the valuation of health states IV, V, and VI. No significant relation was found between this theme and the TTO utilities ($r = -.18$, $p = .23$).

"*After nineteen years I will be fifty. After that age, your quality of life will diminish, so, that is the same as having RA."* (female, 31, V)

"In about 21 years, I will be an old man … a little bit at least. Everything will become more troublesome anyway. You will develop problems by yourself." (male, 22, V)

Anticipated adaptation

A theme that played a role in the TTO scenario was anticipated adaptation. This describes how respondents expected to adapt to the RA health state in due time. Following Ubel et al., we distinguished between psychological and physical adaptation (53). The first concerns the ability to emotionally adapt to a health state, e.g. a shift in personal interests. Respondents mentioned this consideration almost as often for health states I, II, and III (45%) as for health states IV, V, and VI (55%). Psychological anticipated adaptation correlated positively, though not significantly, with utilities ($r = 0.27$, $p = 0.08$).

> "*I am quite hopeful about it ... At the moment when it happens, then you will see how creative you are; how much you can stretch things. This you can see with people. At the moment you think that something is not possible, then they will get the strength from God knows where. And they deal with it, cope with it. If that is human, then also I must have something like that in me.*" (male, 39, I)

> "*It is rather annoying not being able to perform your daily activities; maybe I can become more socially active.*" (male, 45, III)

> "*I have a mother and she is over 85 years of age and she also has rheumatism. Some pain, and problems with performing her daily activities, also she is washed twice a week. So I have a role model, maybe not fair. But I have her as an example. She is happy and gets around quite well.*" (female, 46, IV)

Physical adaptation concerns the physical aspect of adaptation, e.g. getting a wheelchair. Respondents mentioned this aspect almost as often for health states I, II, and III (56%) as for health states IV, V, and VI (44%).

"*I am counting on pain medication.*" (female, 37, III)

"*It is not like you'll be immobile. There will still be aids, wheelchairs and such.*" (male, 23, VI)

"*Severe pain, the first thing that comes to mind is that, nowadays, of course, something can be done about this.*" (male, 39, I)

Only a few respondents indicated that they anticipated not to adapt to the RA health state.

"*Do you think that at a given moment you wouldn't feel that pain anymore if you have it constantly? I think you would always feel that*." (female, 20, I)

Goals

An important theme for respondents was goal aspiration. It influenced how respondents answered TTO questions. Most often, respondents mentioned not being able to achieve goals because of being in a RA health state or living fewer years in good health. A concern about whether goal aspiration was still possible was often expressed as a desire: 1) to either live life in a certain way (preferably in the same way as now), and/or 2) to achieve specific goals that have been set. Regarding the first concern, respondents considered the impact of the RA health state on their way of life, e.g. to take part in sport, to walk, or to be independent. Consequently, this concern took account of the present as well as of the future. The impossibility of achieving goals while living in that health state was mentioned more often for the more severe health states; 72% of the remarks on that theme were made during the valuation of health states I, II and III.

"If I can't do a thing all day, then I don't need to live. Being down all the time. That is not me … I am very attached to my daily activities. I do want to be able to keep on doing stuff." (male, 22, II)

"Well I don't like to be helpless, so to speak, I don't like to need help. Yes, I like being independent. I can't do my daily activities, I can't go to work then. And I have problems with washing or getting dressed, so that really bothers me ... That is not really the life I would like to have." (female, 22, II)

The second concern regarding set goals was mostly considered by respondents in comparison to the number of years in good health offered in the TTO. If the offered number of years was below the number of years needed for goal realization, respondents often preferred the RA health state in which they would still live for their remaining life expectancy. Goals were mentioned, such as raising children, knowing grandchildren, career opportunities, and enjoying a few years of pension. Consequently, this concern regarded the future. The possibility or impossibility of achieving goals in the offered number of years of good health was mentioned equally frequently for all health states. Respondents shifted their focus on different life goals, depending on the future life years involved in the task.

"*Yes, well then I will choose for option 1, because I want to live a long life. I want to become a grandmother etcetera."* (female, 41, V).

"*Still I would choose for a life with wife and children, and above that many years living in good health*." (male, 23, III).

Maximal endurable time

The theme maximal endurable time (MET) was encountered with respect to the more severe health states. MET is the concept that when people live in a health state involving a low quality of life, at first they evaluate it positively, but after a certain period (MET) any additional time spent in that health state is viewed negatively. Consequently, death is preferable after that period (54). Of the remarks made in which respondents demonstrated MET, 72% were made during the valuation of health states I, II, and III.

> "*What is the use of 54 years of living, if you can't do a thing but sit around the house all day?*" (male, 23, III)

> *"It is that what is said here, as it concerns severe pain. That is tough to live with. At a certain point in time, enough is enough."* (male, 28, I)

Table 5.1. Univariate associations between themes and TTO utility for all health states and respondents (RP = reference point).

Theme	**TTO**
RP RA for remaining life expectancy	.41 (p < .01)
RP Optimal health for period x	.14 (p = .35)
RP Current health state	.19 (p = .23)
Psychological anticipated adaptation	.27 (p = .08)
Physical anticipated adaptation	.18 (p = .24)
Impossibility to achieve goal in RA health state	-.24 (p = .12)
Impossibility to achieve goal in period x in optimal health	.13 (p = .40)
Maximal endurable time	-.24 (p = .12)

After univariate analysis (see Table 5.1), a regression analysis was performed in which themes of borderline significance ($p < .15$) were entered simultaneously. The analysis showed that the following themes were relevant to TTO utilities with an explained variance of $R^2 = 37\%$, $p < .000$, (adjusted $R^2 = 33\%$, $p < .000$): Use of the RA health state as reference point ($p < .005$), anticipated psychological adaptation ($p < .01$), and the impossibility to achieve one's goal in the RA health state ($p < .005$). The latter showed a negative relation with TTO utilities. MET did not contribute to the TTO utility anymore.

Labeling

From the qualitative results of this experiment it appeared that the labeling of the health states that we had decided to use had a strong impact, and led to higher utilities than for unlabeled health states. For the eleven people who had someone with RA in their inner circle of acquaintances the average TTO utility was .67, and for the 34 people with no one with RA in their inner circle the average TTO utility was .57 ($p = 0.17$). Many references that were made to the disease (Rheumatoid Arthritis) were related to themes such as anticipated adaptation, and use of pain killers. We therefore compared our respondents' utilities with those of the British and Dutch tariff (see Figure 5.1). Discrepancies were seen for the poor health states, (I: $p < .000$, II, and $p < .005$),[1] in which our TTO utilities were higher than Dutch EQ-5D utilities. Small discrepancies were seen in the other direction for health states V ($p < .01$), and VI ($p < .02$). We could, however, not rule out that our different utilities were due to the life expectancy we used, which was longer than the 10-year time

[1] The EQ-5D provides a 'tariff' of values for EQ-5D health states. Discrepancies were assessed per health state using the relative disutility (RD) between valuations: RD = (1 - mean TTO)/(1 - EQ-5D index). Since the EQ-5D index has a broader utility range (-0.6 to 1.0) than the TTO (0.0 to 1.0), RD between methods is more representative than the absolute difference. The RD should be 1 if respondents scored in the same way as the general population in the tariff study.

frame used in the tariff study. We therefore carried out an additional experiment to test our assumptions that labeling influenced our utilities.

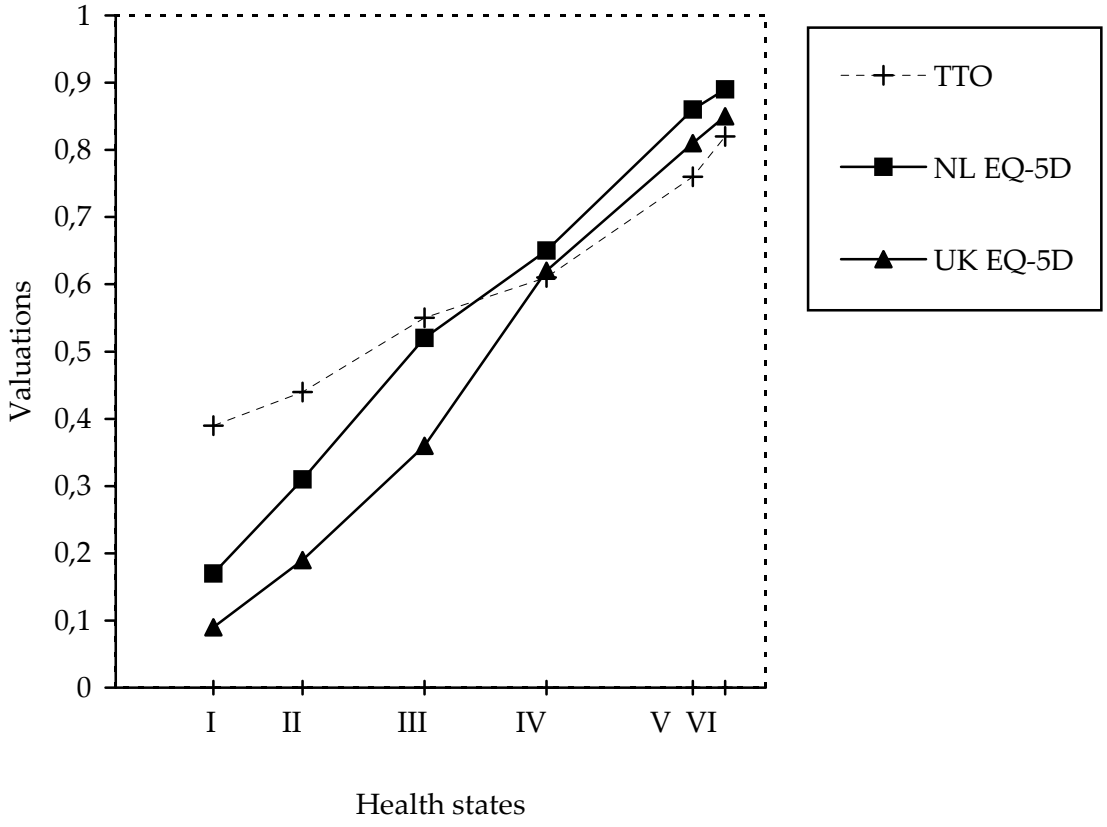

FIGURE 5.1. Mean valuation for each health state per method. The six health states are ranked on the x-axis according to the corresponding mean valuation.

Experiment 2

Methods

Eighty-four of the 112 medical students attending a Medical decision making course at the Leiden University Medical Center participated. Respondents were asked to

perform a TTO (paper version) as a matching task for the six health states used in Experiment 1. The students randomly received one of two versions: health states labeled as rheumatoid arthritis, or unlabeled health states. The remaining life expectancy was set at 10 years to equal the EQ-5D tariff studies. Using a repeated measures ANOVA, with Health states as within and Label as between factor, we assessed the overall effect of labeling. This was followed by independent t-tests to evaluate the effect of labeling for the different health states.

Results

The respondents of the No-Label group consisted of 31 women aged 21 to 46 years (mean age = 24, s.d. = 5) and 16 men aged 21 to 31 years (mean age = 24, s.d. = 3). The respondents of the group Label consisted of 26 women aged 21 to 32 years (mean age = 22, s.d. = 4) and 11 men aged 20 to 26 years (mean age = 26, s.d. = 2).

Table 5.2 shows the mean utilities (and s.d.'s) for the six health states. Remarkable are the differences in standard deviations between the two groups, with the No-label group showing significantly larger s.d.'s ($p < 0.03$) for states I to IV. Overall, the difference between the two groups reached borderline significance (MANOVA, $p = 0.055$). An interaction effect was seen between label and state ($p = 0.10$), with states V and VI not being different between the groups, state IV being significantly different ($p = 0.01$) and states I to III reaching borderline significance ($p \leq 0.10$).

Table 5.2. Mean utility (and s.d.) for labeled and unlabeled health states

	I	**II**	**III**	**IV**	**V**	**VI**
Label	0.56 (.20)	0.57 (.18)	0.67 (.15)	0.75 (.14)	0.85 (.13)	0.94 (.05)
No Label	0.47 (.27)	0.49 (.24)	0.59 (.25)	0.64 (.23)	0.84 (.15)	0.94 (.05)

General discussion

In the first experiment, we found evidence that the impaired health state is indeed most often used as the reference point. According to Bleichrodt, this would lead to higher TTO utilities as a result of loss aversion (6). Our data supports this finding, as the use of this reference point was associated with higher utilities. Additionally, it appeared that the main focus for the TTO in our study lay on length of life rather than on quality of life, suggesting scale compatibility. As a choice task was used, the number of years offered to live in optimal health changed according to the answer given, and it is therefore natural that this answer should receive most attention, as has been observed for the SG in other studies (51;57). We therefore decided not to further code these statements. Bleichrodt argued that scale compatibility would lead to higher utilities when the focus during the task lies more on quantity of life than on quality of life.

We found little evidence for the bias of utility curvature (34). Linearity of the utility for life years is questionable if living for your remaining life expectancy, or a period close to it, is viewed as an unattractive prospect (theme: anticipating problems in old age). This was mentioned sometimes, mostly for the less severe health states, and despite the statement that these years will be spent in optimal health. This theme was

observed in another study as well (70). However, we observed no significant relation with TTO utilities.

The qualitative data showed that respondents incorporated several themes into their decisions. Themes that have been argued in the literature to attribute to the TTO utilities are: maximal endurable time (54), subjective expectations about length of life and quality of life (70;71), family circumstances, and religious beliefs (72). We will discuss themes considered by respondents, and evaluate the possible effects on TTO utilities, and the possible effect of labeling on TTO utilities.

Each person aspires to live to a certain age: the aspiration level (42). Life goals could be a factor strongly involved in the personal setting of the aspiration level. The role of the aspiration level in medical decision making has become more and more acknowledged (29;43;44;47;73). We used the remaining LE as time frame and, therefore, allowed the aspiration level to have a role in TTO valuation, as is the case in genuine situations. Our findings and analysis lead us to argue that respondents considered the aspiration of goals an important theme. The impossibility to realize one's goals while living with RA led to significantly lower utilities.

Maximal endurable time was considered by respondents, especially for the severe health states. The presence of this theme should lead to lower TTO utilities, although the effect did not reach significance in our study. In a regression analysis, the relation with TTO utility was still negative, but the effect was less significant.

In the first experiment, discrepancies between EQ-5D index and TTO existed. For the health states involving a low quality of life, the EQ-5D index provided much lower valuations than the TTO in our study. We hypothesized that labeling of our health

states resulted in this difference, given the importance of anticipated adaptation, and the association with familiarity with RA. A study by Gerard et al. (65) showed a shift in valuation that was partly due to labeling as well. Anticipated adaptation is more likely if a person knows more clearly to what one will be adapting. Psychological anticipated adaptation indeed showed a significant and positive relation with TTO utilities in the first experiment of our study.

However, further evidence was required as the EQ-5D tariff was based on a TTO with a ten year time frame, in contrast to the remaining life expectancy frame that was used in our experiment. In the second experiment that used a more similar TTO method, differences were observed between unlabeled and labeled utilities, although not all reached statistical significance. For two of the six states this absence of significance could be due to a ceiling effect. Remarkable was that for the states in which no ceiling effect was seen, the standard deviations were much larger in the unlabelled condition. As may be expected, labeling makes people think of the same condition. A label provides a more elaborate description, and factors such as experience and familiarity may then be significantly associated with the valuation of the health state (53). Imagination can flow freely without a label, resulting in more variation. Respondents used a broader range of the utility continuum, and, as a result, utilities became more scattered.

Conclusions

We found evidence that the bias loss aversion led to higher TTO utilities. Scale compatibility leads to higher TTO utilities, and we found evidence that this occurs for the TTO. In our experiment, the bias of utility curvature appeared to have no significant effect on TTO utilities. The impossibility to realize goals in the RA health state led to lower utilities, whereas the RA health state as reference point, and psychological anticipated adaptation led to higher utilities. As the reference point most often used by our respondents was the health state to be valued, the bias loss aversion will also most likely operate similarly for patients as well as the general public. It has been argued that patients use less of the utility continuum than the general public when valuing similar health states. Labeling may have a similar effect, but to a lesser extent (74). Patients valuing their own health implicitly include labeling. Consequently, labeling a health state could potentially lower the discrepancies between patient and public estimates of quality of life.

6

Understanding VAS valuations

Qualitative data on the cognitive process

Understanding visual analog scale valuations using qualitative data.

S.M.C. van Osch, A.M. Stiggelbout

Quality of Life Research: 2005; 14: 2171-2175.

Abstract

Eliciting people's values is a central pursuit in health economics. We explored approaches to valuing a health state on a visual analog scale (VAS). Additionally, we examined whether dual processing (an interaction between automatic and controlled information processing) occurred during VAS valuation. In the first experiment, respondents were probed for their approach after valuation on a VAS. After inductive generalization, we grouped the approaches: 1) 'Sort-of' (automatic processing), 2) 'Bisection of line first', 3) 'Numerical expression', and 4) 'Dividing into smaller segments'. In the second experiment, a short questionnaire followed the VAS in which these approaches were systematically assessed, as was *Awareness* of the approach used, *Intention* to re-use the approach the next time (confidence), and *Basis* of the approach. Data showed that the 'Sort-of' approach was used most often, followed by the 'Bisection-first' approach. We argue that dual processing occurs during performance on the VAS. Awareness of the approach used was lower when an intuitive approach was used. A reasoned approach had a higher correlation with confidence. Thus, awareness of approach may improve reliability. Reducing the number of health states to be valued concurrently diminishes the complexity of the task; this may enhance the validity of the VAS.

Introduction

The visual analog scale (VAS) is a widely used valuation technique. Devlin, Hansen and Selai (75) reported an extensive study on the valuation of health states using a VAS. They emphasized the need for research to underpin the cognitive processes underlying the VAS. We have qualitative data that allows us to address this subject. This chapter elaborates on the types of approaches to valuing a health state on a VAS. Additionally, we assessed whether the approaches used are in concordance with dual-processing theory, which states that behavior is determined by the interaction between controlled and automatic processing (76).

Devlin et al. raised several questions; here we would like to address their fourth question, "What types of approach to valuing hypothetical health states using a VAS are detectable?" In their study, the approaches were not systematically elicited, therefore their data cannot answer this question fully (75). Moreover, most approaches mentioned in their study were related to a flawed execution. We believe our data can further elaborate on possible approaches leading to usable valuations. The VAS is a seemingly easy method often used to estimate a person's preference value for health states. Health state valuation involves introspection, evaluation and comparison (77). After this, the valuation is expressed involving several cognitive processes. The VAS is then, basically, a visuospatial motor task.

Cognitive processes can be carried out at two distinct levels with qualitatively different mechanisms. Dual-processing theory states that thoughts, behaviors and feelings result from the interplay of automatic (and implicit) and controlled (and explicit) processing (12). Automatic information processing is ubiquitous and the default mode of processing. The controlled information processing requires controlled

allocation of processing resources (i.e. controlled attention), and, consequently, takes more time (78). In the literature there is a large body of research, for example in the field of social psychology and decision making, indicating that dual processing occurs during a large number of cognitive processes (12;79-81). However, to the best of our knowledge, no research exists that explores whether dual processing is involved in the VAS valuation task.

We carried out two experiments in which respondents were probed for approaches used in the VAS. Possible approaches were explored in the first experiment and were systematically examined in the second. If dual processing is indeed relevant to the VAS, we should find evidence of an implicit and explicit level of thinking. Furthermore, we assume that an explicit approach is more reliable than an implicit approach (1). Therefore, we assessed the basis and awareness of an approach as well as the intention to re-use it. The measurement subsequently employed to assess the valuation is very accurate (to the millimeter). However, the VAS instruction gives no such indication of accuracy to the respondent. Therefore, in the second experiment, we also assessed the effect of adjusting the VAS instruction.

Experiment 1

Methods

Procedure

Healthy respondents were recruited using newspaper advertisements. They were paid € 22.50 for participation in two interviews. A subset of these respondents participated in this experiment. We used a rheumatoid arthritis health state description according

to the EQ-5D system (description = 21321) (25), see Appendix 6A. The valuation for this health state was assessed with a VAS: a 100 mm horizontal line with the anchors 'death' and 'optimal health'. The sixteen respondents were afterwards asked to elaborate on their approach. Qualitative data was taped, and transcribed. Analysis involved initial familiarization with the data through sorting and indexing. Through discussion of emergent approaches during data generation possible approaches were detected. One coder coded all comments.

Results

Seven females (mean age = 38) and nine males (mean age = 37) participated. They were educated to at least high school standard. Qualitative data showed that the majority of the respondents were incapable of specifying why the mark was placed at that precise point, other than that it felt right (57% of all approaches reported, 'Sort-of'-approach), e.g.

> *"So I put it nearer death, but not too much. It feels like the right spot." and "I came to the decision because if you have some problems with walking, that is not too bad, … no problems washing and getting dressed, that is very nice. You are not capable of performing your daily activities, which is very annoying. Then you are bored out of your mind and can't go anywhere … it was more roughly done, a bit of feeling and moving."*

Many respondents started the task by deciding whether the health state was nearer to death or optimal health by bisecting the line (29% of approaches, 'Bisection of line first'-approach), e.g.

"*Well, I go directly to the middle of the line and then I move to the right because one can get accustomed to everything, including pain.*"

Few respondents gave a numerical expression before placement (8% of approaches, 'Numerical expression preceding placement'-approach), e.g.:

"*One provides a sort of value to it, and say, perfect health is a ten, death is zero. Then you provide a value to this health state. It is not death and it is not optimal health. Let's say perfect health is ten and death is zero. … So … that would make the quality of such a health state; … if you translated it to the quality of life in that case … it would be … I'd give it a three on a scale of ought to ten. And that is why I put the mark at that spot."*

In only 3% of the approaches was the VAS divided into smaller segments ('Small segments'-approach) e.g.

"*And that is how I end up at about three-quarters*."

We grouped approaches into four categories through inductive generalization: 1) Sort-of, 2) Bisection of the line first, 3) Numerical expression preceding placement, and 4) Dividing the line into smaller segments. Most respondents used more than one approach simultaneously; mostly, the 'sort-of'-approach was applied in combination with another approach. We labeled the 'sort-of'-approach as an implicit approach for it appeared based at an automatic level of thinking. Whereas the other approaches were labeled as explicit approaches for they appeared based at a controlled level of thinking.

Experiment 2

Methods

Procedure

Fifty-eight healthy medical students attending a course were paid €5 for filling out a questionnaire, part of which related to this experiment. Respondents valued the rheumatoid arthritis health state of experiment 1 by using one of two VAS instructions to which they were randomly assigned (Appendix 6A). Instruction 2 included instruction 1, plus a statement that the number of millimeters from the anchors would be measured to assess the valuation. Possible approaches deduced from experiment 1 were 'Sort-of', 'Bisection first', 'Numerical expression', and 'Smaller segments'. Experiment 1 showed that approaches do not have to be mutually exclusive, hence respondents were asked to indicate on a five-point answering scale the concurrence with their approach ("Does not concur with my approach - Concurs fully with my approach"), separately, for each category. Then the following three questions were answered on a five-point scale:

- (Awareness) "To what extent are you aware of the approach you used to perform this task?" "I am not aware of using an approach - I am aware of using an approach"
- (Basis of approach) "You can perform this task more on the basis of intuition or more on the basis of reasoning. Where do you place your approach?" "Intuitive - Reasoned"
- (Confidence) "Would you probably use the same technique the next time?" "Unlikely to use the same approach – Likely to use the same approach".

A last item inquired whether respondents used an approach not mentioned in the questionnaire. A t-test was performed to assess the effect of Instruction on

questionnaire items. For all respondents, average score per item was assessed, and the (Pearson) correlations between the questionnaire items were assessed.

Results

Twenty-eight respondents were assigned to Instruction 1 (19 females, 9 males, mean age= 21), thirty respondents to Instruction 2 (24 females, 6 males, mean age= 20). No significant effect was seen for Instruction, either for values or questionnaire items. Therefore, we will discuss the results for both instructions.

Respondents most often used the 'Sort-of'-approach (mean = 3.45, s.d. = 1.23), followed by the 'Bisection first'-approach (mean = 2.97, s.d. = 1.52), the 'Numerical expression'-approach (mean = 2.24, s.d. = 1.41), and the 'Small segments'-approach (mean = 2.19, s.d. = 1.41). Significant negative correlations were observed between the 'Sort-of' and 'Bisection first'-approaches ($r = -.27$, $p < 0.05$), and the 'Sort-of' and 'Small segments'-approaches ($r = -.40$, $p < 0.01$). Between the 'Bisection first' and 'Small segments'-approaches, no significant correlation existed ($r = .18$). The ‘Numerical’-approach was not significantly related to another approach.

Most respondents reported the intention to use the same approach the next time they would perform a VAS (mean = 4.28, s.d. = .81). Slightly more respondents were aware of their approach (mean = 3.33, s.d. = 1.03) than unaware. A reasoned approach was used somewhat more frequently (mean = 3.55, s.d. = 1.47) than an intuitive approach.

Table 6.1. Correlations between the questionnaire items for all respondents

	Awareness (Unaware –Aware)	**Basis (Intuitive – Reasoned)**	**Confidence (Intention to re-use approach)**
Awareness	1.00		
Basis	.59**	1.00	
Confidence	.29*	.21	1.00
Sort-of	-.38**	-.62**	-.14
Bisection first	.28*	.18	-.08
Small segments	.27*	.46**	.08
Numerical	.01	.12	-.03

* $p < 0.05$ (2-tailed); ** $p < 0.01$ (2-tailed)

Table 6.1 shows that when a 'Sort-of'-approach was used the valuation was more often intuitive, whereas if respondents used a 'Small segments'-approach, the valuation was more often reasoned. If respondents were more aware of their approach, they used a reasoned approach more frequently than an intuitive approach. Awareness also correlated with the intention to use the same approach the next time (Confidence). Confidence appeared not to be directly related to a specific approach. Merely one respondent reported an approach not described in the questionnaire (First she placed the mark, consequently she assessed whether the health state fitted the position; if it did not, the process was repeated).

Discussion

This study tried to provide a better understanding of cognitive processes underlying the VAS. Our experiments showed four not mutually exclusive approaches to valuing

a health state by using VAS. These approaches were, in order of the respondents' preferences, Sort-of, Bisection of line first, Numerical expression, and Division into small segments. The Numerical and Small segments approaches were used less frequently. This coincides with the (visual) concept of the VAS; otherwise a rating scale should be preferred. Respondents appeared reluctant to numerically express their valuation. Adjusting the task instruction by indicating the way responses were to be used did not encourage respondents to use a more explicit approach. The effect of interest was small to medium; consequently, the power was low (0.06) to medium (0.41).

We found evidence of both automatic and controlled information processing during the VAS within subjects. The combination of an implicit and automatic approach ('Sort-of') and an explicit and controlled approach (other reported approaches in this study) was present in most respondents' answers. This underpins the occurrence of dual processing during valuation of a health state using the VAS. However, to confirm that dual processing occurs during the VAS one should focus on the time aspect of the task performance in relation to the approach used. Controlled information processing requires more time than automatic information processing and, as a consequence, an explicit and reflective approach requires more time than an automatic approach (78). Our study did not include reaction time as a variable, therefore, it does not provide conclusive evidence. It would be interesting to involve this variable, as well as to explore possible approaches with respect to other health status measurement techniques, e.g. standard gamble and time trade-off.

Awareness correlated negatively with the implicit approach. Respondents who were more aware, indicated a stronger willingness to use the same approach the next time. The questionnaire may have, in the less aware respondents, induced the idea that next

time they should give the task more thought. (We cannot assess this, because we did not ask which approach they would use the next time.)

In contrast to the study by Devlin et al., we used a horizontal VAS and asked subjects to value a health state in isolation. Devlin et al. observed that the valuation task used in their study as well as in the EQ-5D, is too complex. Consequently, the cognitive burden was sizable (75). Most approaches reported in their study led to unusable or problematic valuations. Our findings do not show these undesired approaches, either because of our systematic assessment of approaches in comparison to their by-product assessment, or because the VAS used in this chapter was less involved (the health state was valued in isolation). Bleichrodt & Johannesson also argued against valuing several health states simultaneously (82). In lessening the number of health states that are valued concurrently, the complexity of the task is diminished, and possibly the validity of the VAS is enhanced. A disadvantage, however, is that rank ordering of many health states would no longer be possible. Thus, a trade-off has to be made. Rank ordering the health states beforehand as a separate task can further refine the validity of the valuation process.

Both in our study and in the study by Devlin et al., some respondents used a 'numerical' approach. The other approaches we reported were not found in their study. However, detecting these approaches required cognitive interviewing which Devlin et al. did not aim for. Other approaches in their study, leading to usable valuations, are specific for the vertical VAS, and, therefore, were not observed in our study.

Our qualitative analysis revealed several possible approaches that were coded by only one coder. Furthermore, the process of valuing appears to be partly automatic,

therefore, it is possible that more approaches exist than are reported here. Although an argument in favor of reliability is that, in the second experiment, only one respondent indicated using an approach that was not reported in the questionnaire.

We present evidence of dual processing during health state valuation using the VAS. Controlled processing, for example, being aware of an approach, may enhance reliability of the VAS since the use of a similar strategy on two occasions stands a higher chance of producing the same result. Additionally, awareness is related to explicit approaches (such as bisecting the line first, or dividing it into smaller segments). These findings are an argument - for the sake of reliability - in favor of instructing respondents beforehand, in order to determine their approach.

Appendix 6A. VAS instructions

Instruction 1

" How do you value the following rheumatoid arthritis health state?"

Some problems walking about

No problems washing or dressing

Unable to perform usual activities (work, family, leisure)

Moderate pain or discomfort

Not anxious or depressed

We would like you to indicate on this scale how good or bad you find this health state. Please do this by placing a mark on this line. The closer to the left you place the mark, the more you value the health state in the same way that you would value death. The closer to the right you place the mark, the more you value the health state in the same way that you would value optimal health."

Instruction 2

Instruction 1 and

"Afterwards we will assess your valuation for this health state by measuring the number of millimeters, in which death is equivalent to 0 millimeters and optimal health is equivalent to 100 millimeters."

7

The development of the Health-Risk Attitude Scale

The development of the Health-Risk Attitude Scale.

S.M.C. van Osch, A.M. Stiggelbout

Submitted for publication

Abstract

People differ in their attitude towards health risks. This results in differences in preventive health risk behavior and treatment preferences. We developed the health-risk attitude scale (HRAS) in order to assess how persons value their health and manage health risks. The HRAS aims to predict how a person will resolve risky health decisions in the future. Items for the scale were devised mostly on the basis of findings in the literature and through interviews with patients on their treatment preference. The psychometric aspects of the HRAS scale were tested in two studies. To assess construct validity, we used general risk taking scales, a domain specific risk scale, standard gambles, the health locus of control scale and a personality scale. Study 1 describes the construction of the first version of the HRAS and documents the validity and reliability. After factor analysis and reliability analysis the HRAS consists of 13 items. Study 2 describes the validity and reliability of the final version of the HRAS. Relations with other risk scales were positive. In summary, we have developed a short and simple scale assessing risk attitude in a health context. It shows good reliability (both internal and test-retest) and convergent validity.

Introduction

People differ in their attitude towards health risks, and this results in different health risk behavior whether preventive (e.g. diet) or other (e.g. treatment preference) (83-85). There exists a relation between risk attitude and behavior. A person with a positive attitude towards risks will undertake more risk than someone who is risk averse. Differences among individuals in their attitude towards health risks may also be expected to provide valuable information about treatment preferences (33;86;87) and medical decision making (83). There are different ways to assess risk attitude, e.g. a scale or the standard gamble certainty equivalent (CE) method based on expected utility (EU) theory. There are several scales to assess general risk attitude (83;84), but there are no scales to assess health related risk attitude. A CE method involves a time consuming interview to assess a person's health related risk attitude. We developed the health-risk attitude scale (HRAS) in order to assess how persons value their health and manage health risks.

It is difficult to decisively classify individuals as risk-averse or risk-seeking because they behave differently in different situations. Some regard risk attitude as a situation-specific concept (88). Risk attitude can be stable across domains as formulated by Weber et al. (2002). They argued that all persons have a negative attitude towards risk (are risk averse) and that a distinction should be made between risk perception and attitude towards perceived risk. They used perceived risk as a measure of riskiness, and argue that most people are systematically risk averse across domains. (Perceived) risk attitude is inferred by regressing risk taking on risk perception. They argue for the existence of a general risk attitude trait which is obscured by situations and domains that affect risk perception (89). For example, if a person takes risks in the financial domain but not in the social domain, this is due to his/her not perceiving the financial

situation as risky, while perceiving the social situation as risky. EU describes the form of the utility function derived from a series of risk-related questions. According to EU, risk attitude is a trait, and a descriptive label of the utility function. If, however, Weber's argument is valid, then risk attitude is confounded with risk perception.

We developed a health-risk attitude scale instead of a general risk attitude scale. By keeping the scale situation-specific, we circumvent inquiring about a person's risk perception as argued by Weber et al (89). The HRAS only aims to predict how a person will resolve risky health decisions in the future. More knowledge about a person's health risk attitude enables health care providers to better understand their own and patients' treatment preferences. We aimed for a short, and easy to administer scale.

Construct validity of the HRAS was assessed by comparing the scale to other scales. To assess convergent validity we included other risk attitude, risk perception and risk taking scales. We expect the HRAS to show a high correlation with scales that measure risk attitude in the same domain, and a lower correlation with more general risk scales, or risk scales involving other domains as well. We argue that a relation exists between risk perception and risk behavior (89-91). Often risk attitude is inferred from risk behavior. Risky behavior is (at least partly) the result of a positive attitude towards risk together with a low risk perception. Therefore, a negative relation is predicted between risk perception and risk attitude (89).

To assess discriminant validity, we included a scale that measures health locus of control. Previous studies have linked locus of control with health behavior, but not with health risk attitude. People with an internal locus of control (who believe that

their own actions influence their health) are more likely to adopt health-promoting behaviors. We expected a small and negative correlation between risk attitude and internal locus of control. In other words, people with a high internal locus of control, being more aware of their risks, take fewer health risks (89).

We expect a weak relationship between health risk attitude and personality. However, we found no study linking personality and risk attitude. Studies do report that personality plays a role in risk behavior (92;93) and, hence, it may very well be related to risk attitude as well. Personality factors such as extraversion and openness are suggested to have a positive relation with overall risk propensity, while factors such as conscientiousness, agreeableness, and neuroticism have a negative relation (94). With respect to the proposed relation between personality and risky health behavior (e.g. smoking), contradictory results have been found (92).

Study 1 describes the construction of the first version of the HRAS and documents the validity and reliability of the first version. Study 2 describes the validity and reliability of the final version of the HRAS.

Study 1 Reliability and validity of HRAS (first version)

Methods

Scale development

To develop the items for the HRAS, we performed an extensive pilot study and a literature study. Part of the pilot study was an interview of two patients with severe

autoimmune disease. They underwent an experimental treatment (a high dose chemotherapy followed by autologous stem cell transplantation) which is risky compared to conventional treatment (1 - 3% risk of death) (95). The interviews were about patients' motivations and beliefs with respect to their risk attitude and the relation to the treatment chosen. Furthermore, we explored qualitative data from interviews with 103 Dutch patients with resectable rectal cancer about their treatment preferences (96). These patients commented on their treatment as well as on related risks. We not only included items that were related to medical treatment, but also items related to preventive and risky health behavior. We further assumed that people's health risk attitude is related to their attitude to health in general and, therefore, we also included items that incorporated such attitudes. Based on the interviews and the literature, a pilot version of the health-risk attitude scale was constructed. In total 43 respondents of various ages and backgrounds filled out pilot versions of the health risk scale. Based on inter-item correlations and feedback provided by the pilot respondents, we selected 18 items (see Table 7.2). Each item was followed by a seven-point Likert scale on which the extent of agreement could be indicated (1 = totally disagree to 7 = totally agree).

Respondents

There are three samples. Sample A consisted of 26 women (mean age = 28, s.d. = 15) and 19 men (mean age = 34, s.d. = 14). They filled out the questionnaires as part of another experiment on construction of health state utilities (51). Respondents were recruited through newspaper ads and were paid 25 Euros for participation in the entire experiment. Sample B consisted of students from the Leiden University Medical Center attending a course at medical decision making. Questionnaires were distributed and 50 students returned their questionnaires (response rate = 49%). The respondents were 35 women (mean age = 24, s.d. = 5) and 15 men (mean age = 23, s.d.

= 3). They were paid five Euros for returning the scale. Sample C consisted of 58 medical students from the Leiden University Medical Center following a course. Of the 85 questionnaires that were handed out, 58 were returned (response rate = 68%) by 43 women (mean age = 20, s.d. = 3) and 15 men (mean age = 21, s.d. = 3). Students were paid five Euros for participation. In total 147 respondents participated. Due to time constraints, it was not possible for each respondent to fill out all scales of interest (Table 7.1 shows to which samples which instruments were administered).

Construct validity

We computerized seven standard gamble Certainty Equivalents (CE), CE12.5, CE25, CE37.5, CE50, CE62.5, CE75, CE87.5. A CE is a standard gamble for which probabilities are held constant, in our case at .5. The outcomes of the gamble were in healthy life years. The certain outcome is varied until indifference results. The so-called risk parameter (r) is based on the seven CEs, the indifference outcomes. It measures risk attitude according to EU in this study with respect to healthy life years. If $r < 1$ then according to EU a person is risk averse. If $r = 1$, then a person is risk neutral, if $r > 1$ then a person is risk seeking. For more information on CEs, calculating r and the precise procedure used, we refer to Van Osch et al. (34).

The Domain-Specific Risk Attitude Scale (DOSPERT) consists of two scales, one measuring risk taking (deduced from risk behavior) and another measuring risk perception (89). It measures across six domains, i.e. Gambling (4 items), Investment (4 items), Health/Safety (8 items), Recreational (8 items), Ethical (8 items) and Social (8 items). For the risk behavior scale, respondents indicate the likelihood of engaging in 40 different activities on a five-point rating scale ranging from 1 (Extremely unlikely) to 5 (Extremely likely). The higher the score the more risk seeking one behaves. The risk perception scale inquired to what extent these activities were perceived as risky

on a five-point rating scale ranging from 1 (Not at all risky) to 5 (Extremely risky). The higher the score the more an activity is viewed as risky. We translated the DOSPERT into Dutch using a forward-backward procedure. We replaced two items, and reworded five more items to adapt the scale to the Dutch culture. Moreover, in consultation with one of the authors of the original scale, we added, "should the occasion occur" to the existing instruction ("Please indicate the likelihood of engaging in the following activities") of the risk behavior part of the DOSPERT. The Dutch version is available upon request. Appendix 7A shows results of the construct validity and test-retest reliability of the Dutch version of the DOSPERT.

The short Jackson Personality Index (JPI) measures general risk taking (83). It consists of six items on a five-point rating scale ranging from 1 (Strongly disagree) to 5 (Strongly agree). The higher the score, the higher the general risk taking is.

Lion's risk scale (RS) is also a general risk-taking tendency scale (97). It consists of eight items, each followed by a nine-point rating scale ranging from 1 (Strongly disagree) to 9 (Strongly agree). Again the higher the score, the higher the degree of risk taking is.

The Health Locus of Control Scale (H-LOC) consists of three subscales, i.e. Internal orientation, Doctor orientation, and Chance orientation (98;99). Each subscale consists of 6 items. Each item is followed by a six-point scale ranging from 1 (Strongly disagree) to 6 (Strongly agree).

The five-dimensional personality test (5- DPT) measures personality on five different factors: Insensitivity, Extraversion, Neuroticism, Orderliness, and Absorption (100). The scales consists of 100 items on a "yes/no"- scale.

A subset of sample A (10 women, mean age = 36 and 12 men, mean age = 38) completed the HRAS twice within a two week interval to assess test-retest reliability.

Table 7.1. Samples of study 1, number of respondents and scales that were filled out.

Scale (# of items)	Sub A (N = 22)	A (N = 45)	B (N = 50)	C (N = 58)
HRAS (18)	x		x	x
HRAS retest (18)	x			
DOSPERT (80)		x	x	x
DOSPERT retest (80)		x		
CEs (7)		x		
JPI (6)	x		x	x
RS (7)		x	x	x
H-LOC (18)		x	x	x
5-DPT (100)		x		x

Data analysis

Items 2, 5, 6, 7, 11, 13, 14, 16, and 17 of the HRAS needed to be recoded. The higher the score, the more positive one's attitude is towards health risks. Scores could range from 18 – 126. We performed a factor analysis (principal components analysis followed by a varimax rotation) to explore the scale for constructs (possible subscales) recoding items. We assessed internal reliability of the scale by calculating internal consistency (Cronbach's alpha) and test-retest reliability by Intraclass correlation (ICC). To assess construct validity, we calculated correlations between scales (Pearson's). The missing values of the HRAS were examined. Missing values were imputed if more than 50% of the items of the scale was present.

Results

Two of the 147 respondents together had five missing items in the HRAS. Two respondents were unable to fill out the 5-DPT due to time constraints.

Factor analysis

The eighteen items loaded on six factors (eigenvalues > 1). Factor one explained 28% of the variance, and factors two and three only explained 9% each. Factor four, five and six explained 8%, 7% and 6% respectively. Based on the scree plot (see Figure 7.1), we evaluated a one-, two- and three-factor model. In the two-factor model, most items loaded on factor one (Table 7.2) and no general construct could be observed for the items 4, 8, and 11 that loaded on factor two in the two-factor model. The three-factor model did not point to any clearly interpretable factors. We specified a one-factor model (Table 7.2).

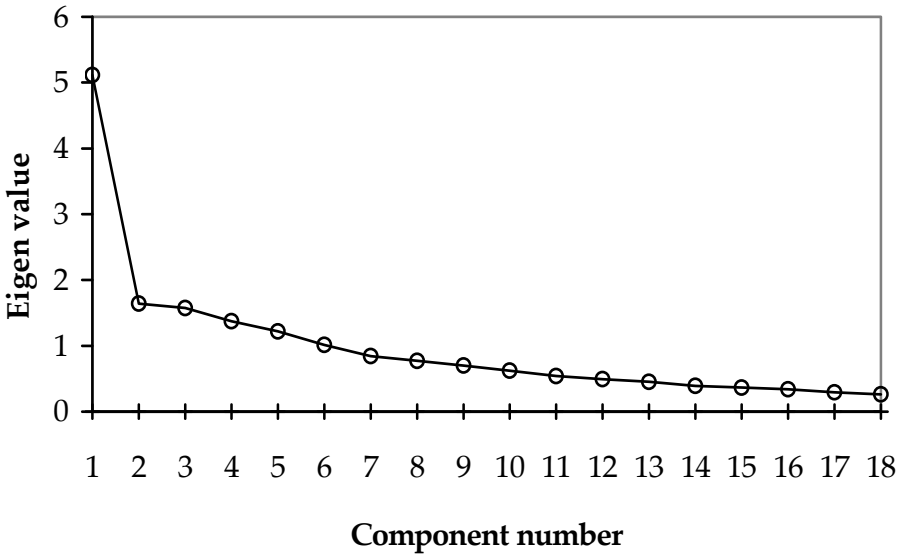

FIGURE 7.1. Scree plot of the factor analysis of the HRAS, first version (18 items).

Table 7.2. Factor analysis of first version HRAS (based on one-factor model). Items with a * are included in the final version. Factor loadings above .30 and below -.30 are depicted.

Item	Factor loading
1. I don't have a problem with taking risks with my health if the benefits are great enough.	.44
2. I think that I take good care of my body.*	.61
3. I don't want to have to consider the consequences for my health with everything that I do every time I do something.*	.49
4. If, due to illness, I could die prematurely, then I would accept a high risk operation which could prevent this from happening.	
5. It is important for me that I organize my life so that I will later enjoy good health.*	.61
6. If it concerns my health, then I see myself as someone who avoids risks.*	.76
7. My health means everything to me.*	.63
8. When I think of an operation, I immediately think of the risk.	
9. When I look back at my past, I think that, in general, I did take risks with my health.*	.59
10. I'm not very fussy about my health.*	.76
11. I would accept risks in undergoing a medical treatment only if I thought there are no reasonable alternatives.	
12. Uncertainty about the consequences of a medical intervention is, in general, part of the game.*	.36
13. Safety first, where my health is concerned.*	.75
14. To enjoy good health now and in the future, I am prepared to forego a lot of things.*	.61
15. People say that I take risks with my health because of my habits.*	.57
16. If the doctor cannot offer me certainty about the possible consequences of a medical intervention, then I would rather not undergo it.*	.36
17. I would never want to have an operation with a high mortality risk no matter what my symptoms are.	.48
18. In general I would estimate that I would not have much of a problem with undergoing a high risk operation.	.45

Reliability

We performed a reliability analysis on items that leaned on the first factor (not including items 4, 8 and 11), based on the analysis items 1 and 17 were removed. Internal consistency of all 13 items now included was .84 (see Appendix 7B for final version of the HRAS). Removal of any other items would not improve internal consistency. Test-retest reliability (based on the 13 items) was computed for the subset of sample A, and the ICC was high at .85. The mean score for all respondents was 62.4, the standard deviation was 13.1, and the median was 62.

Validity

Table 7.3 shows correlations between the 13-item HRAS and other (sub)scales. The HRAS showed a positive but non-significant correlation with the risk parameter based on CEs. The risk parameter only showed a positive correlation with Lion's risk scale ($r = .34$, $p < .05$). As predicted, the HRAS showed mostly positive and moderate to low correlations with the DOSPERT Behavior subscales, except for the correlation with the subscale Health (.50) which was good. The correlation between the HRAS and the DOSPERT perception scale was negative for all scales. The highest correlation was again observed for the subscale Health. The general risk scales (RS and JPI) showed reasonably good correlations with the HRAS.

A negative, and almost significant correlation ($r = -.20$, $p = .07$) was observed with the subscale Doctor orientation of the H-LOC. For the 5-DPT, there was a negative and high correlation ($r = -.40$, $p < .000$) between the Orderliness (e.g. to be neatly or to prefer it methodologically arranged) subscale and the HRAS.

Table 7.3. Correlations between HRAS, first version, and other scales.

Scales	HRAS
CE	.23
D-Per Soc	-.07
D-Per Recr	-.27**
D-Per Inv	-.27**
D-Per Gamb	-.01
D-Per Eth	-.19
D-Per Health	-.39**
D-Beh Soc	-.07
D-Beh Recr	.21*
D-Beh Inv	.34**
D-Beh Gamb	.39**
D-Beh Eth	.37**
D-Beh Health	.50**
RS	.51**
JPI	.45**
HLOC I	.04
HLOC D	-.20
HLOC C	.07
5-DPT E	-.02
5-DPT N	-.16
5-DPT A	.01
5-DPT I	.19
5-DPT O	-.40**

**Correlation is significant at the 0.01 level, *correlation is significant at the 0.05 level.

Discussion

The purpose of this study was to evaluate the psychometric properties of the HRAS. Factor analysis did not identify subscales in the HRAS. After factor analysis, reliability analysis showed that five items could be removed to improve reliability. The HRAS showed high reliability, both test-retest reliability and internal consistency.

Construct validity was assessed with various scales. The convergent validity was reasonable to good with all correlations in the expected direction. The highest correlations were observed with the RS and the subscale Health of the DOSPERT. Lower correlations were observed with other subscales of the DOSPERT. This indicated good convergent validity. As expected, correlations with the HRAS were negative for the DOSPERT perception scale and positive for the DOSPERT behavior scale. This agrees with the findings by Weber et al. (89). We had expected a higher correlation between the HRAS and the CEs. The CEs showed a significant correlation with a more general risk tendency scale. The HRAS correlated negatively with the personality dimension Orderliness.

Study 2 Validity and reliability of the final version of the HRAS

Study 2 served to validate the 13-item version of the HRAS, both in students and in a more general sample.

Methods

Respondents

Sample D consisted of medical students of the Leiden University Medical Center attending a course on scientific education. Of the 100 questionnaires, 43 were returned (response rate = 43%). The respondents consisted of 36 women (mean age = 19, s.d. = 1) and 9 men (mean age = 20, s.d. = 2), two respondent won a cinema ticket after participation. Sample E consisted of employees from a training and consultancy firm and from the Dutch organization for council of clients from retirement and nursing homes. Two co-workers handed out 110 questionnaires and 89 respondents responded (response rate = 81%). Respondents consisted of 58 women (mean age = 43 years, s.d. = 15) and 31 men (mean age = 45 years, s.d. = 18). They received no reward for participation. All respondents completed the final version of the HRAS, the DOSPERT, RS, and JPI. Items 1, 3, 4, 6, 8, 9, and 10 of the HRAS need to be recoded. The higher the score, the higher the health-related risk attitude.

Results

Three respondents had 6 missing items from the HRAS. The mean score of all respondents was 42.3, the standard deviation was 8.8, and the median was 42. No significant difference in the HRAS was observed between samples.

Reliability

The internal consistency of the HRAS was reasonable at .71. If item 5 was removed it increased to .75. In the final version, only three of the 13 items involve risk related to a medical treatment, one of which is item 5. Therefore, we choose to maintain item 5 as

part of the HRAS. The removal of any other item resulted in a decrease of the reliability.

Construct validity

Table 7.4 shows correlations between the HRAS and the DOSPERT, RS, and JPI. All correlations were in the expected direction. The strongest correlation with the DOSPERT subscales was again with the health/safety scale. The RS and JPI showed positive relations with the HRAS.

Table 7.4. Pearson correlations between HRAS, final version, and other scales.

Scales	HRAS
DR-B Inv	.03
DR-B Gam	.09
DR-B Health	.33**
DR-B Recr	.22*
DR-B Eth	.13
DR-B Soc	-.05
DR-P Inv	-.14
DR-P Gam	-.22*
DR-P Health	-.37**
DR-P Recr	-.32**
DR-P Eth	-.17
DR-P Soc	-.05
RS	.49**
JPI	.25**

**Correlation is significant at the 0.01 level, *correlation is significant at the 0.05 level.

General discussion

The main aim of this chapter was to assess the reliability and validity of a health related risk attitude measure, the Health-Risk Attitude Scale. On the basis of the data showed we argue that it assesses health related risk attitude. The first results from using the HRAS indicate that it is a short and, hence, tractable scale that appears to adequately measure health related risk attitude. The response rate to our study was reasonable.

Internal and test-retest reliability are good. A further indication that the HRAS has a good construct validity is the good convergent validity; the correlation with other risk scales is significant. Moreover, the highest correlation with the DOSPERT by Weber et al. was observed with the health/safety subscale. The Health/safety subscale of the DOSPERT does not incorporate health related risk attitude with respect to medical treatments, and can therefore not substitute for the HRAS. The items of the Health/safety subscale involve mostly preventive health behavior, e.g. the use of sunscreen or the engaging in unprotected sex. The HRAS correlates equally well with the risk taking and risk perception DOSPERT Health/safety subscale. It may be possible that the apparent difference in willingness to take health risks may actually be mediated by differences in the perception of such risks. Future research should focus on the relation between the perceived riskiness of a medical treatment (or health behavior, e.g. smoking) and the willingness to undergo a treatment (or perform health behavior).

The HRAS did not correlate with the CEs. The CEs in our experiment involved future healthy life years (i.e. length of life), and not quality of life. In other words, they measure risk attitude in relation to time preference. Possibly, if the CE involved more

of the qualitative aspect of health, we would have observed a better relation. Moreover, the CEs showed no significant relation to the DOSPERT subscale Health/safety, but it showed the strongest relation to general risk taking tendency. The relationship between risk attitude and time preference needs to be explored further (101).

There exists no relation between health locus of control and health risk attitude. We nevertheless found studies linking health locus of control and health (preventive) behavior. Risk attitude and health locus of control appear to be two different, unrelated constructs that influence health risk behavior to a different extent. There is a negative relation between personality trait Orderliness and health risk attitude. Orderliness is a variant of the Big-five dimension Conscientiousness (100;102). In another study, a negative relation was also observed between Conscientiousness and health risk propensity, but this relation was not significant. Our findings support the natural reasoning that a person who would take care to do things carefully and correctly is more inclined to view health risk taking negatively.

In summary, we have developed a short and simple scale assessing risk attitude in a medical context. It shows good reliability and convergent validity. The next step is to test the scale in studies assessing medical decision making in high risk contexts, such as organ transplants.

Appendix 7A. Reliability and validity of the Dutch DOSPERT based on studies 1 and 2.

Test-retest reliability of the DOSPERT behavior scale varied per subscale. ICC for the subscale Social (.47) was low, correlations were moderate for the subscales Health (.52), Gambling (.55), Investment (.62), and Ethics (.68). Correlation was high for the subscale Recreational (.89). For the DOSPERT perception scale, the ICCs were overall somewhat lower. They were low for the subscales Health (.33) and Gambling (.43), and moderate for Investment (.56), Recreational (.60), Ethics (.66) and Social (.69).The DOSPERT subscales Health/safety and Recreational showed either the highest or the most frequently a correlation with other scales, see Table 7A.

Table 7A. The DOSPERT (D in table) consists of two scales, i.e. behavior (B) and perception (P), each is made up of six subscales: Investment (Inv), Gambling (Gam), Health/Safety (Health), Recreational (Recr), Ethics (Eth), and Social (Soc). Correlations between Dutch DOSPERT and other scales are depicted. The short sensation seeking scale (SSS) measures risk taking behavior involving thrill and sensation (103). The Revised Life Orientation Test (LOT) measures dispositional optimism (104). For other scale abbreviations, number of respondents, see Methods section of Study 1.

Scales	**D-B Inv**	**D-B Gam**	**D-B Hea**	**D-B Recr**	**D-B Eth**	**D-B Soc**	**D-P Inv**	**D-P Gam**	**D-P Hea**	**D-P Recr**	**D-P Eth**	**D-P Soc**
RS	.17**	.19**	.43**	.45**	.22**	.16**	-.27**	-.15*	-.35**	-.35**	-.25**	-.13*
JPI	.18**	.21**	.29**	.35**	.23**	.15*	-.28**	-.12	-.28**	-.28**	-.23**	-.17**
SSS	.21**	.11	.23**	.50**	.20*	.02	-.14	.04	-.12	-.41**	-.12	-.06
LOT	.08	-.10	.12	.25*	-.09	.07	.06	.05	-.13	-.25*	-.11	-.09
HLOC I	.10	.14	-.02	.14	.17	.16	.02	.00	.01	.00	.16	.08
HLOC D	.19	.01	-.10	-.22*	-.21*	.06	-.03	-.07	.10	.18	.17	.08
HLOC C	.08	-.06	.09	-.23*	.00	-.17	.07	.10	.01	.11	-.08	.04

**Correlation is significant at the 0.01 level, *correlation is significant at the 0.05 level.

Appendix 7B. Final version

H-RAS (Van Osch & Stiggelbout)

On the following pages you will find several statements. Read through each statement. Please circle the number which best reflects your opinion. There are no right or wrong answers. We are interested in your opinion.

For example:

I am a morning person

Totally disagree	1	2	3	4	5	6	7	Totally agree

By putting a circle around number 4, you would indicate that you neither agree nor disagree with this statement.

If you totally disagree with the statement, then put a circle around number 1.

If you disagree to a great extent with the statement, then put a circle around 2.

If you slightly disagree with the statement, then put a circle around 3.

If you slightly agree with the statement, then put a circle around 5.

If you agree to a great extent with the statement, then put a circle around 6.

If you totally agree with the statement, then put a circle around 7.

1. I think I take good care of my body.
2. I don't want to have to consider the consequences for my health in everything that I do.
3. It is important to me that I organize my life so that I will later enjoy good health.
4. If it concerns my health, then I see myself as someone who avoids risks.
5. Uncertainty about the consequences of a medical intervention is, in general, part of the game.
6. My health means everything to me.
7. When I look back at my past, I think that, in general, I did take risks with my health.
8. If the doctor cannot offer me certainty about the possible consequences of a medical intervention, then I would rather not undergo it.
9. Safety first, where my health is concerned.
10. To enjoy good health now and in the future, I am prepared to forego a lot.
11. People say that I take risks with my health because of my habits.
12. I'm not very fussy about my health.
13. In general I would estimate that I would not have much of a problem with undergoing a high risk operation.

8

Summary and Conclusions

This thesis presents the results of our examination of the construction of health state utilities according to expected utility (EU) and (cumulative) prospect theory (PT). PT extends EU and encapsulates its biases and inconsistencies. The aim of our study is to investigate decision making from both a normative and a descriptive point of view and to improve the measurement of utility in medicine through findings from non-expected utility theory, i.c. PT. This chapter summarizes and draws a number of conclusions from the preceding chapters.

Chapter 1 provided an outline of the thesis. In particular, the distinction between normative, prescriptive, and descriptive models was discussed. Two commonly used methods to measure utilities are the standard gamble (SG) and time trade-off (TTO). The SG is based on normative expected utility (EU) arguments. However, there is much empirical evidence that expected utility is not descriptively valid and that its violations generate upward biases in SG utilities. Among other things, these violations are described in the descriptive model PT. Although lacking the theoretical foundations of the SG, the TTO has emerged as the most frequently used method due to better feasibility, higher discriminative power and better face validity. No empirical research has been carried out yet on the biases inherent in this method.

Chapter 2 provided new insights into existing correction methods for the biases, advanced in the economic literature, and to test them in the medical domain. TTO utilities have been suggested to be pushed downwards by the bias of utility curvature, and upwards by loss aversion and scale compatibility. Thus it has been suggested that TTO biases may neutralize each other. The biases of probability weighting and loss aversion generate an upward effect for SG utilities. The effect of the bias scale compatibility on SG utilities was as yet unknown (see chapter 4 also). In PT, outcomes are described as gains or losses (positive or negative deviations) relative to a (neutral)

reference point (32). Therefore, the reference point dictates how to correct SG utilities for probability weighting and loss aversion. However, little is known about the psychology behind the location of the reference point, for which PT does not include a hypothesis. For the health domain, there is no direct evidence concerning the location of the reference point.

In the data presented, the gains-corrected SG showed the best convergence with TTO scores. However, this chapter provided arguments suggesting that the TTO utilities were biased upwards, rather than having balanced biases. Utility curvature was absent at the average level, and, as a result, correcting for utility curvature had little effect at this level. Moreover, the TTO scores were higher than the theoretically most preferred correction of the SG, the mixed correction. These findings suggest once more that uncorrected SG scores, which were higher than TTO scores, are too high. Further, it is likely that uncorrected TTO scores are also too high. It remains an empirical question, however, whether the theoretically preferred mixed correction of the SG is the appropriate correction. This was evaluated in chapter 4. Qualitative data from both chapters 4 and 5 provided further insights into these findings, and will be summarized and discussed below.

Chapter 3 presented information about the reference point in certainty equivalent (CE) standard gambles. We assessed the outcome in the CE that seemed closest to, or seemed to include, the reference point. Additionally, the psychology behind the reference point was explored. Qualitative data were combined with quantitative data to provide evidence of the reference point in life-year CE gambles and to explore the psychology behind the reference point. On the basis of direct and indirect evidence we concluded that the offered CE most often served as the reference point and, therefore, that a life-year CE gamble is most likely to be perceived as a mixed gamble. This finding can have important consequences for economic evaluations, since for mixed

gambles the corrected utilities differ from the uncorrected utilities (11;34). Framing and goals influenced the perception of the reference point. Goals also caused an interaction between the time horizon and the attention paid to an outcome. We argued that motivational constructs should be incorporated into research on the reference point. Because patient preferences are involved in treatment decisions, it is preferable to incorporate a patient's goals into treatment decision making.

Chapter 4 further examined the effect of biases on probability equivalent SG utilities using qualitative data on the reference point and focus of attention. To assess the effect of scale compatibility, correlations were assessed between focus of attention and mean utility. The certain outcome, i.e. the health state to be valued, most often served as the reference point. The SG was then most likely to be perceived as a mixed gamble. Additionally, goals were mostly mentioned with respect to this outcome. These findings can have important consequences for cost-effectiveness analyses, such as correcting preference-based utilities for loss aversion and probability weighting with the appropriate mixed correction formula (11;34). Additionally, we observed that respondents mostly focused on the low outcome of the gamble. Consequently, scale compatibility would still have led to upward biases in the SG utilities.

Chapter 5 provided evidence based on qualitative data that indeed the impaired health state is (most often) used as the reference point in the TTO, as argued by Bleichrodt (4). The use of the impaired health state as reference point was associated with higher utilities as a result of loss aversion. Additionally, it appeared that the main focus for the TTO in our study lay on quantity of life as opposed to quality of life, which would lead to higher utilities through scale compatibility (6). We found little evidence of the bias of utility curvature (34). Moreover, we observed no significant relation between the bias of utility curvature and TTO utilities. It has been argued that

patients use less of the utility continuum than the general public when valuing similar health states. We observed that labeling may have a similar effect, but to a lesser extent.

Chapter 6 dealt with cognitive processes underlying the VAS. Strictly speaking, the VAS is not a preference-based measure, and therefore does not provide utilities. It is nevertheless often substituted for SG or TTO, for reasons of feasibility. We therefore evaluated this measure as well. Our experiments showed four not mutually exclusive approaches to valuing a health state by using a VAS. These approaches were, in order of the respondents' preferences, Sort-of, Bisection of line first, Numerical expression, and Division into small segments. The Numerical and Small segments approaches were used less frequently. This coincides with the (visual) concept of the VAS; otherwise a rating scale should be preferred. Respondents appeared to be reluctant to express their valuation numerically. Adjusting the task instruction by indicating the way responses were to be used did not encourage respondents to use a more explicit approach. Next, we examined whether dual processing (an interaction between automatic and controlled information processing) occurred during VAS valuation. We presented evidence of dual processing during health state valuation using the VAS. Controlled processing, for example, being aware of an approach, may enhance reliability of the VAS since the use of a similar strategy on two occasions stands a higher chance of producing the same result. Additionally, awareness is related to explicit approaches (such as bisecting the line first, or dividing it into smaller segments), and also to having confidence in the answer. These findings are an argument in favor of instructing respondents beforehand to determine their approach, to improve reliability.

Chapter 7 dealt with the measurement of health related risk attitude. People differ in their attitude towards health risks. This results in different preventive health risk behavior and treatment preferences. In medical decision making, until now mostly Certainty Equivalent standard gambles have been used to assess risk attitude. Elicitation is a complex task, fraught with biases, as explained above. We therefore developed the health-risk attitude scale (HRAS) in order to assess how persons value their health and manage health risks, and refer to this as health risk attitude. The HRAS aims to predict how a person will resolve risky health decisions in the future. Items for the scale were devised mostly through a literature study and through interviews with patients about their treatment preferences. The psychometric aspects of this instrument were tested in two studies. To assess construct validity, we used general risk taking scales, a domain specific risk scale, standard gambles, the health locus of control scale, and a personality scale. Study 1 described the construction of the first version of the HRAS and documented the validity and reliability. On the basis of a factor analysis and a reliability analysis, the HRAS was reduced to 13 items. Study 2 described the validity and reliability of the final version of the HRAS. Relations with other risk scales were positive. In summary, we have developed a short and simple scale assessing risk attitude in the health context. It showed good reliability (both internal and test-retest) and convergent validity.

Most people working in the medical domain seem to understand intuitively that the SG is biased upward, because most people resort to the TTO to assess utilities. The introduction of PT into medical decision making has provided an explanation for the intuitive gut feelings that the SG utilities generally were much too high. The TTO has removed the standard gamble from its golden throne and has gained power due to its better descriptive abilities. On the basis of qualitative and quantitative data, we have found that scale compatibility generates an upward effect on SG utilities. The

qualitative data provided at the end of this thesis on the SG method argued for a mixed-corrected SG. All biases effective in the SG method push SG utilities upwards.

The data concerning the SG reference point further underpins the argument stated in chapter 2 that TTO utilities are still too high, because TTO utilities were higher than the mixed-corrected SG utilities. Additional support for this argument was provided by the qualitative data on TTO utilities, because the counterbalancing bias of utility curvature was infrequently found. This may have been due to our sample, though, because other studies did find this curvature. On the other hand, Stalmeier (105) has pleaded for the use of uncorrected TTOs for utility curvature. The other two biases were observed, as well as the upward generating effect thereof on TTO utilities. We propose to use the mixed-correction formula to correct SG utilities (derived with ping-pong elicitation method) proposed by Bleichrodt, Pinto and Wakker (4). No correction is available as yet for scale compatibility, and more research is clearly needed here. TTO utilities appeared to be too high. However, there is no correction method for scale compatibility today. Awareness of the total effect of the biases affecting TTO utilities is all that we offer now. More research is required for this issue.

The two generic models of cognitive function (dual processing) that were observed during the VAS, appeared also to be present during the TTO and SG. These are the intuitive mode in which judgments and decisions are made and the reasoned (more controlled) mode in which these are made deliberate but in a slower pace. Both appeared to play a role during the elicitation. The distinction between intuition and reasoning has been a topic of interest (106;107). Emotions and beliefs play a big role in decision making for which EU holds no room. EU, like 'traditional theories of economic decision-making' assumes that humans are fundamentally rational

creatures. However, humans are reproducibly irrational in a number of characteristic ways.

Utility cannot be divorced from emotion. It has even been found that the framing effect was specifically associated with neural activity in a key emotion center in the human brain. This finding also highlights the importance of incorporating emotional processes within models of decision making (108). In decision analysis, it is commonly assumed that the right normative model for decision under uncertainty is expected utility. Some implicitly assume the EU model to have no deviations and biases by using uncorrected standard gamble utilities. We can conclude that this is incorrect. A further question that occurs is: Can one say that the expected utility theory is correct if so much deviations and biases occur? Bleichrodt, Pinto and Wakker argue that biases and inconsistencies are not to be interpreted as irrationalities on the client's part, but are the result of measurement techniques (4).

However, what is so bad about irrationality? In the case of individual decision making there may be nothing wrong with that. Emotion is often regarded as the antithesis of reason. However emotion (e.g. fear) can often be thought of as a systematic response to observed facts. It is safe to argue that emotion is complex. Emotions and rationality have a variable and complex connection, sometimes they strengthen one and other, sometimes they are each others enemy. I can have emotions that are grounded in judgments with which I do not agree. I can even be angry that I feel such unfounded emotions, such as with respect to my fear of spiders. However, my reason can justifiably be told introspectively that it is being unreasonable. We cannot deny that emotions play a role in the way that people arrive at decisions with respect to their physical well being (109). It is possible to formalize these emotions, and these should be incorporated in descriptive decision models. The difficult question that remains is

whether emotions should be incorporated into a prescriptive model. Based on the findings in this thesis we can acknowledge that goals in life and emotions play an important role in decision making, but based on these findings we cannot answer the question posed above. However, a theory of decision making that completely disregards feelings such as anticipated regret and loss aversion is not only descriptively invalid, it also results in prescriptions that do not maximize the utility of outcomes as they are actually experienced (106). If emotions are a standard part of the cognitive process that is decision making, then how can one prescribe how to decide without them?

Considering the available decision models, it is not necessarily true that when utilities are corrected for biases, that these biases should not exist. Rather, these biases are not incorporated into the existing models. On the other hand, all elements that are relevant to decision making cannot be expected to be incorporated in one model, as only a limited number of parameters can be used in order to maintain tractability of a model. As argued before, for the time being, we should then be satisfied with what we have, for decisions must be made now (4). The first step is to have a growing awareness of the effect of biases on utilities in the field of medical decision making. The next step is to actually correct for them. A third step is to further formalize the role of emotions in the decision models. But bear in mind that, it still cannot 'incorporate the wind caused by the wings of a butterfly'.

Samenvatting

Weersvoorspellingen betreffen de toestand van de atmosfeer op een bepaald tijdstip en op een bepaalde locatie. Voorspellingen volgen een model. Maar geen enkel model kan alle variabelen insluiten die relevant zijn. *Zo kan een weersvoorspelling niet de wind meenemen die wordt veroorzaakt door de vleugels van een vlinder, ook als dat de oorzaak is van een latere orkaan.*

Het is gebruikelijk om gezondheidswinst uit te drukken in termen van kwantiteit en kwaliteit van leven. Een utiliteit is een maat voor de waardering van de kwaliteit van leven. Dit proefschrift behandelt de samenstelling van gezondheidsutiliteiten volgens twee besliskundige modellen, namelijk verwachte nutstheorie (expected utility (EU)) en prospect theorie (PT). EU is een normatieve theorie die voorschrijft hoe rationele beslissingen genomen zouden moeten worden volgens iemands eigen preferenties. Echter er zijn dusdanig sterke en systematische empirische afwijkingen van EU gevonden, dat men voor descriptieve doeleinden is gaan zoeken naar verbeteringen. PT is zo een verbetering. Het is een puur descriptieve theorie en het heeft geen normatieve claims. PT breidt EU uit en omschrijft de fouten en inconsistenties ervan. Dit proefschrift analyseert beslissingen zowel met behulp van het normatieve model, EU, als met behulp van het descriptieve model, PT. Doel is meten van utiliteiten binnen de gezondheidszorg te verbeteren met behulp van descriptieve bevindingen van PT.

Hoofdstuk 1 beschreef de opzet van dit proefschrift. Vooral het onderscheid tussen normatieve, prescriptieve en descriptieve modellen werd behandeld. De drie meest gebruikte methodes om utiliteiten te schatten werden besproken. Dit zijn de standard gamble (SG), de time trade-off (TTO) en de Visueel analoge schaal (VAS).

De SG heeft als enige methode een goede theoretische basis en vindt zijn oorsprong in EU. Voor de SG wordt de respondent gevraagd te beoordelen welke kans op overlijden hij of zij bereid is te accepteren om in optimale gezondheid te verkeren in plaats van een te waarderen gezondheidstoestand (de "zekere uitkomst"). Het probleem is dat mensen niet rationeel beslissen, alhoewel de onderliggende EU theorie dat wel aanneemt. De consequentie is dat de SG utiliteit meetfouten bevat. De fouten (zoals veroorzaakt door kansweging als door verliesangst) en de mogelijke correctiemethodes, worden onder andere omschreven in PT. Het referentiepunt speelt in PT een belangrijke rol. Het is het uitgangspunt waarnaar men het handelen of denken kan richten. Echter PT biedt voor medische toepassingen geen concrete aanwijzingen voor wat respondenten als referentiepunt ervaren.

Voor de TTO wordt de respondent gevraagd te beoordelen hoeveel jaren leven in optimale gezondheid men maximaal wil inleveren om niet in de te waarderen gezondheidstoestand (van minder goede kwaliteit) voor de resterende levensverwachting te leven. De TTO wordt ook verstoord door een aantal meetfouten, onder andere als gevolg van de incorrecte aanname dat we ieder toekomstig jaar evenveel waarderen. In de literatuur wordt gesuggereerd dat de TTO utiliteit de ware utiliteit het dichtst benadert omdat deze het minst last heeft van meetfouten.

De VAS is een lijn van tien centimeter met als uiteinden dood en optimale gezondheid. De respondent wordt gevraagd aan te geven ter hoogte van welk punt t.o.v. de eindpunten de te waarderen gezondheidstoestand zich bevindt wat betreft zijn waarde. De VAS geeft over het algemeen lagere waarderingen dan de SG en TTO. De grootste kritiek op de VAS is dat de methode geen theoretische basis heeft.

In hoofdstuk 2 presenteerden we nieuwe inzichten in de correctiemethodes voor de meetfouten. De correctiemethodes worden besproken in de economische literatuur, maar veelal genegeerd in de medische literatuur. Dat is betreurenswaardig, omdat het corrigeren van meetfouten leidt tot betere schattingen van utiliteiten. SG utiliteiten worden beïnvloed door de meetfouten verliesangst, kansweging en schaal-compatibiliteit. TTO utiliteiten worden beïnvloed door de meetfouten verliesangst, utiliteitskromming en schaalcompatibiliteit. Verliesangst wil zeggen dat we gevoeliger zijn voor een bepaald verlies dan voor evenzoveel winst. Kansweging wil zeggen dat mensen kansen wegen, en daarbij overmatig gevoelig zijn voor kansen dichtbij onmogelijkheid (0%) en zekerheid (100%). Utiliteitskromming betreft het afnemende marginale nut van levensjaren. Schaalcompatibiliteit wil zeggen dat wanneer een attribuut van een mogelijke uitkomst van een beslissing (bijvoorbeeld levensduur) veel overeenkomsten vertoont met de respons methode (bijv. zoals het geval in de TTO, waar een antwoord gevraagd wordt in termen van levensjaren), dit attribuut meer gewicht krijgt in een beslissing.

In PT worden uitkomsten beschreven als winst of verlies ten opzichte van een (neutraal) referentiepunt. Iets wordt als verlies of winst gepercipieerd vanuit een neutraal referentiepunt. Als je € 50 hebt gewonnen, dan lijkt dat positief. Als je echter hoort dat je anders € 60 gewonnen had, dan voelt het opeens als negatief. Dit is het gevolg van een verschillend referentiepunt. In het eerste geval ervaart men een uitkomst als winst, en in het tweede geval ervaart men dezelfde uitkomst als een verlies. Het referentiepunt bepaalt hoe SG utiliteiten gecorrigeerd moeten worden voor kansweging en verliesangst. Na correctie zijn dan mogelijk: (a) de voor winst-gecorrigeerde SG utiliteit; (b) de voor verlies-gecorrigeerde SG utiliteit; of (c) de gemengde gecorrigeerde SG utiliteit. De laatste is een half winst, half verlies perceptie van de SG methode en wordt gezien als de theoretische favoriet. Dit wil zeggen dat de

zekere uitkomst, oftewel de te waarderen uitkomst, het referentiepunt was. Veelal leiden de meetfouten tot een te hoge schatting van de SG utiliteit. Het effect van schaalcompatibiliteit op de SG utiliteit was bij de start van dit onderzoek nog onbekend (zie hoofdstuk 4). Er is in de literatuur gesuggereerd dat de meetfouten in de TTO elkaar zouden neutraliseren.

De data lieten zien dat de voor winst-gecorrigeerde SG utiliteit de beste overeenkomst liet zien met de TTO utiliteit. Echter, een onverwachte uitkomst van ons onderzoek was dat de TTO voornamelijk last had van opwaartse meetfouten en dus toch een te hoge schatting gaf van utiliteiten. De meetfout utiliteitskromming had geen invloed op TTO utiliteit. Dit had volgens de literatuur de tegenhanger moeten zijn van de andere twee meetfouten die beide een opwaarts effect hebben op TTO utiliteiten. Daarnaast was de TTO hoger dan de theoretisch favoriete correctie van de SG, de gemengde correctie methode. We toonden verder aan dat zoals al vaker werd gevonden, de ongecorrigeerde SG inderdaad te hoge utiliteiten geeft. Dit kan belangrijke consequenties hebben voor kosteneffectiviteitanalyses. Hoofdstukken 3, 4 en 5 voorzien in een verder inzicht van deze bevindingen met behulp van kwalitatieve gegevens.

In hoofdstuk 3 toonden we met behulp van kwalitatieve data dat respondenten meestal de zekere uitkomst als referentiepunt percipieerden in de SG certainty equivalent (CE) methode. Dat wil zeggen dat de CE wordt gepercipieerd als zijnde deels winst en deels verlies. Dit heeft consequenties voor het gebruik van deze methode omdat de gecorrigeerde CE utiliteit verschilt van de ongecorrigeerde. Verder bleek dat het realiseren van gevormde doelen (zoals voor kinderen zorgen, promotie maken en genieten van het pensioen) voor mensen erg belangrijk is en dat dit eveneens de perceptie van het referentiepunt beïnvloedde. Tevens beïnvloedden

doelen de hoeveelheid aandacht die werd geschonken aan een uitkomst en daarmee de utiliteit. Op basis van deze gegevens argumenteerden we dat motivaties zoals persoonlijke doelstellingen van een patiënt een grotere rol zouden moeten spelen in de behandelkeuze van patiënten.

De kwalitatieve data uit hoofdstuk 4 lieten zien dat de te waarderen gezondheidstoestand in de SG meestal fungeert als referentiepunt. Dat wil zeggen dat de SG wordt gepercipieerd als zijnde gemengd. Persoonlijke doelen worden ook meestal gerelateerd aan deze uitkomst. Voorts vonden we indirect bewijs dat de meetfout schaalcompatibiliteit een opwaarts effect heeft op de SG. Van de andere twee meetfouten was reeds bekend dat ze de SG utiliteit omhoog drukken. Wederom vinden we dus bewijs dat de ongecorrigeerde SG utiliteit substantieel te hoog is.

In het vijfde hoofdstuk verzamelden we kwalitatieve data met betrekking tot de TTO. De data lieten zien dat, zoals verwacht, de gezondheidstoestand die wordt gewaardeerd als referentiepunt fungeerde. Dit leidt tot een hogere utiliteit als gevolg van de meetfout verliesangst. Ook werd meer aandacht geschonken aan lengte van leven dan aan kwaliteit van leven, hetgeen leidt tot een hogere utiliteit als gevolg van de meetfout schaalcompatibiliteit. De kwalitatieve data toonde niet de aanwezigheid van de meetfout utiliteitskromming. We beschreven verder onder andere het mogelijke effect van het geven van een label aan de te waarderen gezondheidstoestand. Het lijkt erop dat het gebruiken van een label, zoals reumatoïde artritis, ertoe leidt dat er een minder groot deel van het utiliteitscontinuüm wordt gebruikt. Hierbij lijkt anticiperende adaptatie een rol te spelen.

In hoofdstuk 6 inventariseerden we mogelijke strategieën die men hanteert bij het invullen van de VAS, om zo de onderliggende cognitieve processen te onderzoeken.

Vervolgens werd systematisch nagegaan in welke mate de strategieën werden gehanteerd. Onze experimenten lieten zien dat er vier mogelijke strategieën zijn, in de volgende voorkeursvolgorde: “Zo ongeveer”, “Halveren van de lijn”, “Numerieke uitdrukking” en “Verdeling in kleinere delen”. Indien men vooraf een extra instructie kreeg waarin de doelstelling van het onderzoek werd beschreven, werd men zich niet bewuster van de gebruikte strategie. Vervolgens bleek dat er waarschijnlijk duale informatieverwerkingsprocessen voorkomen tijdens het waarderen op de VAS. Duale informatieverwerkingsprocessen zijn een interactie tussen automatische en gecontroleerde informatieverwerking. Een bekend voorbeeld is het fietsen naar het werk. Dit gaat normaal bijna vanzelf (automatische informatieverwerking). Er is geen of maar weinig aandacht nodig. Echter, als de weg naar het werk is opgebroken, dan moet er een nieuwe route worden bedacht (gecontroleerde informatieverwerking). Hiervoor is wel aandacht nodig. Gecontroleerde processen zoals het zich bewust zijn van de strategie,verbetert mogelijk de betrouwbaarheid van de VAS, omdat het gebruik van eenzelfde strategie een volgende keer eerder hetzelfde resultaat zal geven. We vonden dat bewustzijn van strategie gerelateerd is aan een meer expliciete strategie (zoals eerst de lijn halveren), maar ook aan vertrouwen in het antwoord. Op basis van deze bevindingen beargumenteerden we dat het aanbeveling zou verdienen om respondenten vooraf te instrueren hun strategie te bepalen, bijvoorbeeld tijdens een oefenronde.

In hoofdstuk 7 introduceerden we een risicovragenlijst. Deze lijst meet gezondheidsgerelateerde risicoattitude. Mensen verschillen in hun gezondheidsgerelateerde risico attitude. Dit resulteert in verschillend gezondheidsgedrag, en in verschillende keuzes om een medische behandeling te ondergaan. In de medische besliskunde gebruikt men meestal de SG (of CE) om risico attitude te bepalen. Dit is echter een veelomvattende taak, met meetfouten, zoals

hierboven uitgelegd. De gezondheidsgerelateerde risicovragenlijst (HRAS) probeert te bepalen hoe een persoon haar/zijn gezondheid waardeert en omgaat met gezondheidsgerelateerde risico's. Het uiteindelijke doel is om te voorspellen hoe een persoon met riskante gezondheidssituaties zal omgaan. Items voor de schaal werden ontwikkeld met behulp van literatuurstudie en door interviews met patiënten over hun voorkeur voor een riskante medische behandeling. We testten de psychometrische eigenschappen in twee studies. Om de construct validiteit te bepalen gebruikten we algemene risico schalen, een domein specifieke risicoschaal, SG's, een beheersingsoriëntatieschaal, en een persoonlijkheidsvragenlijst. In studie 1 wordt de constructie van de HRAS omschreven en worden de validiteit en betrouwbaarheid bepaald. Op basis van een factor analyse en een betrouwbaarheidsanalyse werd de HRAS gereduceerd van 18 items tot 13 items. In studie 2 beschreven we de validiteit en betrouwbaarheid van de uiteindelijke versie van de HRAS. Relaties met andere risicoschalen waren positief. De HRAS liet een goede betrouwbaarheid zien (zowel intern als test-hertest) en een goede convergente validiteit. Het lijkt erop dat de HRAS een korte en simpele vragenlijst is die gezondheidsgerelateerde risico attitude meet.

De SG moet worden gecorrigeerd voor de opwaartse effecten van de meetfouten voor zover daar een correctiemethode voor bestaat. De zekere uitkomst fungeerde veelal als referentiepunt, dus de SG moet gecorrigeerd worden met de gemengde correctiemethode zoals voorgesteld door Bleichrodt, Pinto en Wakker (4). Er is nog geen correctiemethode beschikbaar voor schaalcompatibiliteit.

De enige neerwaartse fout in de TTO, utiliteitskromming, wordt zowel in de kwantitatieve data als kwalitatieve data niet gevonden. Er werd wel kwalitatief bewijs gevonden voor de andere meetfouten. Daarnaast bleek de gemengde gecorrigeerde SG utiliteit nog altijd lager dan de TTO. De TTO lijkt dus te lijden aan voornamelijk

opwaartse meetfouten. Omdat er nog geen correctiemethode bestaat voor verliesangst en schaalcompatibiliteit, is het enige dat we nu kunnen doen ons bewust zijn van de meetfouten die op de TTO werken.

De twee informatieverwerkingsmechanismen die gebruikt worden bij het waarderen met de VAS lijken ook aanwezig te zijn bij het waarderen met de TTO en SG. Er werd zowel op intuïtieve wijze als op meer beredeneerde wijze ingegaan op de TTO en SG. Het onderscheid tussen intuïtie (meer berustend op emoties) en redeneren is een onderwerp dat veelvuldig in de belangstelling staat. Het blijkt dat emoties en overtuigingen een rol spelen in het nemen van beslissingen en dus ook in het waarderen van kwaliteit van leven. Echter zowel emoties als overtuigingen maken geen onderdeel uit van EU. EU heeft, net als andere traditionele theorieën van economische besliskunde, als aanname dat mensen rationele wezens zijn. Echter mensen zijn op soms karakteristieke wijze irrationele wezens.

Een utiliteit kan niet worden gescheiden van emotie. Het framing effect duidt erop dat mensen zich laten beïnvloeden door hoe informatie wordt aangeboden. Het framing effect is speciaal geassocieerd met neurale activiteit in het emotionele centrum van het brein. Dit benadrukt de bevindingen dat emotionele processen een plaats moeten hebben in besliskundige modellen. In besliskundige analyses wordt meestal aangenomen dat het juiste normatieve model voor beslissingen in onzekerheid de EU is. Soms neemt men dit impliciet aan door niet te corrigeren voor meetfouten. We kunnen concluderen dat dat niet correct is. Een verder vraag is: Kunnen we zeggen dat EU correct is als er zoveel fouten en afwijkingen optreden? Bleichrodt, Pinto en Wakker stellen dat de inconsistenties niet het gevolg zijn van irrationaliteit van de persoon maar van gebrekkigheden van de meetmethode.

Maar wat is er zo erg aan irrationaliteit? In de individuele besliskunde is er mogelijk niets mis mee. Emotie wordt vaak gezien als de antithese van rede. Daarentegen kan emotie vaak worden gezien als een systematische reactie op geobserveerde feiten. Emoties spelen een rol in hoe beslissingen worden genomen. Het is misschien mogelijk deze emoties te formaliseren en deze te verwerken in een beschrijvend besliskundig model. Het lastige is het beantwoorden van de vraag of emoties een plaats hebben in een prescriptief model. Op basis van de literatuur en bevindingen uit dit proefschrift kunnen we stellen dat emoties en doelstellingen een rol spelen in de besliskunde, maar we kunnen niet de bovenstaande vraag uit de morele besliskunde beantwoorden. Een theorie die gevoelens zoals geanticipeerde spijt en verliesangst negeert is geen goede beschrijvende theorie. Daarnaast resulteert het in voorschriften die niet de utiliteit van uitkomsten maximaliseren zoals ze eigenlijk wel worden ervaren. Als emoties een standaard onderdeel zijn van het cognitieve proces, betreffende beslissingen, hoe kunnen we dan voorschrijven om te beslissen zonder hen erbij te betrekken?

Het is niet zo dat wanneer men corrigeert voor meetfouten, dat deze fouten niet mogen bestaan. Maar dat punten die aan de meetfouten ten grondslag liggen niet zijn opgenomen in het model. Aan de andere kant, niet alle elementen die relevant zijn voor besliskunde kunnen in een model worden verwerkt, en er zit een limiet aan het aantal parameters dat we kunnen gebruiken in een model om het hanteerbaar te houden. Voorlopig moeten we werken met wat we hebben omdat beslissingen nu genomen moeten worden. De eerste stap die we maken is een groter bewustzijn van het effect van meetfouten op de utiliteiten binnen de medische besliskunde. De volgende stap is om ervoor te corrigeren. Een derde stap is de rol van emoties verder te formaliseren in besliskundige modellen. Maar wees ervan bewust dat het model "*niet de wind kan meenemen die wordt veroorzaakt door de vleugels van een vlinder*."

References

(1) Payne JW, Bettman JR, Schkade DA. Measuring constructed preferences: Towards a building code. Journal of Risk and Uncertainty 1999; 19(1-3):243-270.

(2) Kahneman D. A perspective on judgment and choice - Mapping bounded rationality. American Psychologist 2003; 58(9):697-720.

(3) Von Neumann J, Morgenstern O. Theory of games and Economic Behavior. Princeton, NJ: Princeton University Press, 1947.

(4) Bleichrodt H, Pinto JL, Wakker PP. Making descriptive use of prospect theory to improve the prescriptive use of expected utility. Management Science 2001; 47(11):1498-1514.

(5) Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. Econometrica 1979; 47:263-291.

(6) Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Economics 2002; 11(5):447-456.

(7) Hershey JC, Schoemaker PJH. Probability versus certainty equivalence methods in utility measurement: Are they equivalent? Management Science 1985; 31:1213-1231.

(8) Drummond MF, Stoddart GL, Torrance GW. Methods for the Economic Evaluation of Health Care Programmes. Oxford, GB: Oxford University Press, 1990.

(9) Sox HC, Blatt MA, Higgins MC, Marton KI. Medical decision making. Stoneham, MA: Butterworth-Heinemann, 1988.

(10) Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. Health Services Research 1972; 7:118-133.

(11) Tversky A, Kahneman D. Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty 1992; 5:297-323.

(12) Barrett LF, Tugade MM, Engle RW. Individual differences in working memory capacity and dual-process theories of the mind. Psychological Bulletin 2004; 130(4):553-573.

(13) Delquie P. Inconsistent trade-offs between attributes - New evidence in preference assessment biases. Management Science 1993; 39(11):1382-1395.

(14) Miyamoto J, Eraker SA. Parameter estimates for a QALY utility model. Medical Decision Making 1985; 5(2):191-213.

(15) 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. Journal of Clinical Epidemiology 1995; 48(10):1207-1214.

(16) Pliskin JS, Shepard DS, Weinstein MC. Utility functions for life years. Health Status Utility Functions 1979;207-224.

(17) Bleichrodt H, Pinto JL. The validity of QALYs under nonexpected utility. The Economic Journal 2005; 115:533-550.

(18) Abdellaoui M. Parameter-free elicitation of utility and probability weighting functions. Management Science 2000; 46(11):1497-1512.

(19) Bleichrodt H, Pinto JL. A parameter-free elicitation of the probability weighting function. Management Science 2000; 46(11):1485-1496.

(20) Gonzalez R, Wu G. On the shape of the probability weighting function. Cognitive Psychology 1999; 38(1):129-166.

(21) Robinson A, Loomes G, Jones-Lee M. Visual Analog Scales, Standard Gambles, and Relative risk aversion. Medical Decision Making 2001; 21:17-27.

(22) Tversky A, Sattath S, Slovic P. Contingent Weighting in Judgment and Choice. Psychological Review 1988; 95(3):371-384.

(23) Bleichrodt H, Pinto JL. Loss aversion and scale compatibility in two-attribute trade-offs. Journal of Mathematical Psychology 2002; 46(3):315-337.

(24) de Jong Z, Munneke M, Zwinderman AH, Kroon HM, Jansen A, Ronday KH 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 and Rheumatism 2003; 48(9):2415-2424.

(25) Dolan P. Modelling valuations for EuroQol health states. Medical Care 1997; 35:1095-1108.

(26) Ci3-250. USA: Skim Software Division Rotterdam, Sawtooth Software, Inc, 2000.

(27) Lenert LA, Cher DJ, Goldstein MK, Bergen MR, Garber A. The effect of search procedures on utility elicitations. Medical Decision Making 1998; 18(1):76-83.

(28) Overlevingstafels 2001. Centraal Bureau voor de Statistiek . 2003. 1-6-2003.

(29) Verhoef LCG, de Haan AFD, van Daal WAJ. Risk attitude in gambles with years of life: empirical support for prospect theory. Medical Decision Making 1994; 14:194-200.

(30) 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. Medical Decision Making 1994; 14:82-90.

(31) Bleichrodt H, Pinto JL, Abellan JM. A consistency test of the time trade-off. Journal of Health Economics 2003; 22:1037-1052.

(32) Tversky A, Kahneman D. The framing of decisions and the psychology of choice. Science 1981; 211:453-458.

(33) Prosser LA, Kuntz KM, Bar-Or A, Weinstein MC. The relationship between risk attitude and treatment choice in patients with relapsing-remitting multiple sclerosis. Medical Decision Making 2002; 22(6):506-513.

(34) van Osch SMC, Wakker PP, van den Hout WB, Stiggelbout AM. Correcting biases in standard gamble and time tradeoff utilities. Medical Decision Making 2004; 24:511-517.

(35) Schmidt U. Reference dependence in cumulative prospect theory. Journal of Mathematical Psychology 2003; 47(2):122-131.

(36) Miyamoto J, Eraker SA. Parametric models of the utility of survival duration - Tests of axioms in a generic utility framework. Organizational Behavior and Human Decision Processes 1989; 44(2):166-202.

(37) Nease RF. Risk attitudes in gambles involving length of life: aspirations, variations, and ruminations. Medical Decision Making 1994; 14:201-203.

(38) Lopes LL. Reasons and resources: The human side of risk taking. In: Bell NJ, Bell RW, editors. Adolescent Risk Taking. Newbury Park, CA: Sage, 1993: 29-54.

(39) Schneider SL, Lopes LL. Reflection in preferences under risk: Who and when may suggest why. Journal of Experimental Psychology: Human Perception and Performance 1986; 12(4):535-548.

(40) Harte JM, Westenberg MRM, Vansomeren M. Process Models of Decision-Making. Acta Psychologica 1994; 87(2-3):95-120.

(41) QSR Nud ist Vivo. Australia: QSR International Pty. Ltd., 2000.

(42) Miyamoto J, Eraker SA. A multiplicative model of the utility of survival duration and health quality. Journal of Experimental Psychology-General 1988; 117(1):3-20.

(43) King LA, Richards JH, Stemmerich E. Daily goals, life goals, and worst fears: Means, ends, and subjective well-being. Journal of Personality 1998; 66(5):713-744.

(44) Heath C, Larrick RP, Wu G. Goals as reference points. Cognitive Psychology 1999; 38(1):79-109.

(45) Lopes LL. Re-modelling risk aversion: A comparison of Bernouillian and rank dependent value approaches. In: Von Furstenberg GM, editor. Acting under Uncertainty: Multidisciplinary Conceptions. Boston: Kluwer Academic Publisher, 1990: 267-299.

(46) Lopes LL. Between hope and fear: the psychology of risk. Advances in Experimental Social Psychology 1987; 20:255-295.

(47) Nair KPS. Life goals: the concept and its relevance to rehabilitation. Clinical Rehabilitation 2003; 17(2):192-202.

(48) van den Hout WB, van den Brink M, Stiggelbout AM, van de Velde CJH, Kievit J. Cost-effectiveness analysis of colorectal cancer treatments. European Journal of Cancer 2002; 38(7):953-963.

(49) Quiggin J. A theory of anticipated utility. Journal of Economic Behavior & Organization 1982; 3:323-343.

(50) Robinson A, Loomes G, Jones-Lee M. Visual analog scales, Standard gambles, and relative risk aversion. Medical Decision Making 2001; 21:17-27.

(51) van Osch SMC, Stiggelbout AM, van den Hout WB. Exploring the reference point in Prospect Theory: Gambles for length of life. Medical Decision Making 2006; 26:338-346.

(52) de Jong Z, Munneke M, Jansen LM, Ronday K, van Schaardenburg DJ, Brand R et al. Differences between participants and nonparticipants in an exercise trial for adults with rheumatoid arthritis. Arthritis & Rheumatism-Arthritis Care & Research 2004; 51(4):593-600.

(53) Ubel PA, Loewenstein G, Jepson C. Whose quality of life? A commentary exploring discrepancies between health state evaluations of patients and the general public. Quality of Life Research 2003; 12(6):599-607.

(54) Sutherland HJ, LlewellynThomas HA, Boyd NF, Till JE. Attitudes toward quality of survival: The concept of "maximal endurable time". Medical Decision Making 1982; 2:299-309.

(55) Stalmeier PFM, Bezembinder TGG, Unic IJ. Proportional heuristics in time tradeoff and conjoint measurement. Medical Decision Making 1996; 16(1):36-44.

(56) Wakker PP, Stiggelbout AM. Explaining distortions in utility elicitation through the rank-dependent model for risky choices. Medical Decision Making 1995; 15:180-186.

(57) Baker R, Robinson A. Responses to standard gambles: are preferences 'well constructed'? Health Economics 2004; 13(1):37-48.

(58) Lopes LL. Between Hope and Fear - the Psychology of Risk. Advances in Experimental Social Psychology 1987; 20:255-295.

(59) Horvath P, Zuckerman M. Sensation Seeking, Risk Appraisal, and Risky Behavior. Personality and Individual Differences 1993; 14(1):41-52.

(60) Lopes LL. Risk and Distributional Inequality. Journal of Experimental Psychology-Human Perception and Performance 1984; 10(4):465-485.

(61) Lion R, Meertens RM. Security or opportunity: the influence of risk-taking tendency on risk information preference. Journal of Risk Research 2005; 8(4):283-294.

(62) Damschroder L, Zikmund-Fischer B, Ubel PA. The impact of considering adaptation in health state valuation. Social Science & Medicine 2005; 61(2):267-277.

(63) Dolan P, Gudex C, Kind P, Williams A. Valuing health states: A comparison of methods. Journal of Health Economics 1996; 15(2):209-231.

(64) Dolders MG, Zeegers MP, Groot W, Ament A. A meta-analysis demonstrates no significant differences between patient and population preferences. Journal of Clinical Epidemiology 2006; 59(7):653-664.

(65) Gerard K, Dobson M, Hall J. Framing and labeling effects in health descriptions: Quality adjusted life years for treatment of breast cancer. Journal of Clinical Epidemiology 1993; 46(1):77-84.

(66) Hurst NP, Jobanputra P, Hunter M, Lambert M, Lochhead A, Brown H. Validity of EuroQol - A Generic Health-Status Instrument in Patients with Rheumatoid-Arthritis. British Journal of Rheumatology 1994; 33(7):655-662.

(67) Nord E. Methods for Quality Adjustment of Life Years. Social Science & Medicine 1992; 34(5):559-569.

(68) Paterson C. Seeking the patient's perspective: A qualitative assessment of EuroQol, COOP-WONCA charts and MYMOP. Quality of Life Research 2004; 13(5):871-881.

(69) van Osch SMC, Stiggelbout AM. Understanding VAS valuations: Qualitative data on the cognitive process. Quality of Life Research 2005; 14:2171-2175.

(70) van Nooten F, Brouwer W. The influence of subjective expectations about length and quality of life on time trade-off answers. Health Economics 2004; 13(8):819-823.

(71) Martin AJ, Glasziou PP, Simes RJ, Lumley T. A comparison of standard gamble, time trade-off, and adjusted time trade-off scores. International Journal of Technology Assessment in Health Care 2000; 16(1):137-147.

(72) De Wit GA, Busschbach JJV, De Charro FT. Sensitivity and perspective in the valuation of health status: Whose values count? Health Economics 2000; 9(2):109-126.

(73) Treadwell JR, Lenert LA. Health values and prospect theory. Medical Decision Making 1999; 19(3):344-352.

(74) Wolfe F, Hawley DJ. Measurement of the quality of life in rheumatic disorders using the EuroQol. British Journal of Rheumatology 1997; 36(7):786-793.

(75) Devlin NJ, Hansen P, Selai C. Understanding health state valuations: A qualitative analysis of respondents' comments. Quality of Life Research 2004; 13:1265-1277.

(76) Pashler H, Johnston JC, Ruthruff E. Attention and performance. Annual Review of Psychology 2001; 52:629-651.

(77) Robinson A, Dolan P, Williams A. Valuing health status using VAS and TTO: What lies behind the numbers? Social Science & Medicine 1997; 45(8):1289-1297.

(78) Jansma JM, Ramsey NF, Slagter HA, Kahn RS. Functional anatomical correlates of controlled and automatic processing. Journal of Cognitive Neuroscience 2001; 13(6):730-743.

(79) Sun R, Slusarz P, Terry C. The interaction of the explicit and the implicit in skill learning: A dual-process approach. Psychological Review 2005; 112(1):159-192.

(80) Reber AS. Implicit Learning and Tacit Knowledge. Journal of Experimental Psychology-General 1989; 118(3):219-235.

(81) Winquist JR, Larson JR. Information pooling: When it impacts group decision making. Journal of Personality and Social Psychology 1998; 74(2):371-377.

(82) Bleichrodt H, Johannesson M. An experimental test of a theoretical foundation for rating-scale valuations. Medical Decision Making 1997; 17(2):208-216.

(83) Pearson SD, Goldman L, Orav EJ, Guadagnoli E, Garcia TB, Johnson PA et al. Triage decisions for emergency department. Patients with chest pain: Do physicians' risk attitudes make the difference. Journal of General Internal Medicine 1995;557-564.

(84) Lion R, Meertens RM. Seeking information about a risky medicine: Effects of risk-taking tendency and accountability. Journal of Applied Social Psychology 2001; 31(4):778-795.

(85) Reventlow S, Hvas AC, Tulinius C. "In really great danger..." The concept of risk in general practice. Scandinavian Journal of Primary Health Care 2001; 19(2):71-75.

(86) Gaskin DJ, Kong J, Meropol NJ, Yabroff KR, Weaver C, Schulman KA. Treatment choices by seriously Ill patients: The health stock risk adjustment model. Medical Decision Making 1998; 18(1):84-94.

(87) Fraenkel L, Bogardus ST, Wittink DR. Risk-attitude and patient treatment preferences. Lupus 2003; 12(5):370-376.

(88) Slovic P, Fischhoff B, Lichtenstein S. Risky Assumptions. Psychology Today 1980; 14(1):44-48.

(89) Weber EU, Blais AR, Betz NE. A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. Journal of Behavioral Decision Making 2002; 15(4):263-290.

(90) Cooper A, Woo C, Dunkelberg W. Entrepreneurs' perceived chances for success. Journal of Business Venturing 1988; 3:97-108.

(91) March JG, Shapira Z. Managerial perspectives on risk and risk taking. Management Science 1987; 33:1404-1418.

(92) Vollrath M, Torgersen S. Who takes health risks? A probe into eight personality types. Personality and Individual Differences 2002; 32(7):1185-1197.

(93) Eysenck HJ. Biological dimensions of personality. In: Pervin LA, editor. Handbook of personality: Theory and Research. New York: Guilford press, 1990: 244-276.

(94) Nicholson N, Soane E, Fenton-O'Creevy M, Willman P. Personality and domain-specific risk taking. Journal of Risk Research 2005; 8(2):157-176.

(95) Verburg R, Flierman R, Sont J, Ponchel F, van Dreunen L, Levahrt N et al. The Outcome of Intensive Immunosuppression and Autologous Stemcell Transplantation in Patients with Severe Rheumatoid Arthritis is Associated with the Composition of Synovial T Cell Infiltration. Annals of the Rheumatic Diseases 2005; 64(10):1397-1405.

(96) van den Brink M, van den Hout WB, Jansen SJT, Kievit J, van de Velde CJH, Stiggelbout AM. Assessing costs, quality of life and preferences in patients diagnosed with rectal cancer: Telephone or face-to-face follow-up interviews. Submitted for publication.

(97) Lion R, Meertens RM. Security or opportunity: the influence of risk-taking tendency on risk information preference. Journal of Risk Research 2005; 8(4):283-294.

(98) Wallston KA, Wallston BS, Devellis R. Development of Multidimensional Health Locus of Control (Mhlc) Scales. Health Education Monographs 1978; 6(2):160-170.

(99) Halfens RJG, Philipsen H. Een gezondheidsspecifieke beheersingsorientatielijst. Validiteit en betrouwbaarheid van de MHLC. Tijdschrift voor Sociale Gezondheidszorg 1988; 66:399-403.

(100) van Kampen D. Orderliness as a major dimension of personality: from 3DPT to 4DPT. European Journal of Personality 1997; 11(3):211-242.

(101) Chapman G. Risk attitude and time discounting. Medical Decision Making 1997; 17(3):355-356.

(102) Costa PT, Mccrae RR. 4 Ways 5 Factors Are Basic. Personality and Individual Differences 1992; 13(6):653-665.

(103) Madsen DB, Das AK, Bogen I, Grossman EE. A Short Sensation-Seeking Scale. Psychological Reports 1987; 60(3):1179-1184.

(104) Scheier MF, Carver CS, Bridges MW. Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem) - A Re-evaluation of the Life Orientation Test. Journal of Personality and Social Psychology 1994; 67(6):1063-1078.

(105) Stalmeier PFM, Bezembinder TGG, Unic IJ. Proportional heuristics in time trade-off and conjoint measurement. Medical Decision Making 1996; 16(1):36-44.

(106) Kahneman D. A perspective on judgment and choice: Mapping bounded rationality. American Psychologist 2003; 58(9):697-720.

(107) Epstein S. Integration of the cognitive and psychodynamic unconscious. American Psychologist 1994; 49:709-724.

(108) De Martino B, Kumaran D, Seymour B, Dolan RJ. Frames, biases and rational decision-making in the human brain. Science 2006; 313:5787.

(109) Ubel PA. Emotions, decisions, and the limits of rationality: Symposium introduction. Medical Decision Making 2006; 25:95-96.

Curriculum vitae

Sylvie M.C. van Osch werd geboren in 1976 in Bladel. In 1994 haalde ze haar VWO-diploma aan het St. Jans Lyceum te 's-Hertogenbosch. Vervolgens heeft ze een jaar Bouwkunde gestudeerd aan de Technische Universiteit Eindhoven. In 1995 startte ze aan de opleiding psychologie aan de Radboud Universiteit Nijmegen. In het kader van haar studie heeft ze een half jaar aan de Washington Universiteit te St. Louis, Missouri (VS) onderzoek gedaan naar het leren van motorische sequenties. Ze behaalde haar doctoraal examen in neuro- en revalidatiepsychologie in 2000. In datzelfde jaar startte ze als onderzoeker in opleiding op het project "The construction of health state utilities" bij de afdeling Medische Besliskunde van het Leids Universitair Medisch Centrum. Het onderzoek werd gefinancierd door Zon-MW. De resultaten van dit onderzoek worden in dit proefschrift beschreven. In 2005 kreeg zij voor één van haar artikelen (hoofdstuk 2) de MTA prijs 2005 uitgereikt van de Nederlandse Vereniging voor Technology Assessment in de Geneeskunde. Sinds december 2005 werkt ze als klinisch neuropsycholoog bij de Woonzorggroep Wilgaerden te Hoorn.