Applications of utility theory in the economic evaluation of health care

Toepassingen van nutstheorie in de economische evaluatie van gezondheidszorg

Proefschrift

Ter verkrijging van de graad van doctor aan de Erasmus Universiteit Rotterdam op gezag van de rector magnificus prof. dr. P.W.C. Akkermans M.A. en volgens het besluit van het college voor promoties.

De openbare verdediging zal plaatsvinden op donderdag 18 januari 1996 om 16.00 uur

door

Han Bleichrodt
geboren te Almelo

Promotiecommissie

Promotores:	prof. dr E.K.A. van Doorslaer prof. dr P.P. Wakker
Overige leden:	prof. dr ir J.D.F. Habbema prof. dr G.C. Loomes prof. dr F.F.H. Rutten

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Bleichrodt, Han

Applications of utility theory in the economic evaluation of health care/
Han Bleichrodt - [S.l.:s.n.]. - Ill.
Thesis Erasmus University Rotterdam.
- With ref. - With summary in Dutch.
ISBN 90-9009081-9
NUGI 681
Subject headings: health economics/utility theory/decision under uncertainty/quality of life

Cover: Machiel Crielaard, Utrecht
Printed by: Ridderprint, Ridderkerk

To cure you they must kill you,
the sword of Damocles that hangs above your head.

Lou Reed

Contents

1

Introduction

1.1 QALY based decision making

This thesis studies the applicability of quality-adjusted life years (QALYs) and other utility based outcome measures in medical decision making and health economics. The main conclusion will be that utility based measures are more useful to model health related behaviour than has commonly been thought. To illustrate the applicability of utility based measures, I start with an example.

Consider the following choice situation. An individual must make a comparison between two treatments for a severe skin disease. The first treatment improves the skin disease immediately into a light form of skin disease. The second treatment does not lead to an improvement in the first two years, but after these two years it removes the skin disease. Both treatments are effective for 10 years. That is, treatment 1 leads to 10 years with light skin disease, treatment 2 leads to 2 years with severe skin disease and 8 years without skin disease. The costs of the two treatments have been assessed and have been expressed in monetary units. Which of the two treatment options should be chosen? One possibility is to compare the costs of the two treatment options and to choose the option that is cheapest. However, this criterion, which can be referred to as cost minimization, ignores the fact that the two options produce different outcomes and are therefore not directly comparable. To enable the comparison of both costs and benefits across the two options a way must be found to express the health benefits generated by the two options in a common unit. Given that the costs of the two options are expressed in monetary units, the most obvious measurement unit for the benefits seems to be money. An analysis in which both costs and benefits are expressed in monetary units is referred to as cost benefit analysis. The advantage of cost benefit analysis is that it has a well established foundation in the theory of welfare economics. The monetary value of the health benefits associated with the two options can be assessed by asking individuals how much money they are *willing to pay* (WTP) to

avoid a deterioration of their health from no skin disease to severe skin disease and from light skin disease to severe skin disease respectively. These willingness to pay questions establish the monetary value of the health improvements associated with the two treatment options. Alternatively, the monetary value of the health improvements could be assessed by asking individuals how much money they are *willing to accept* (WTA) for a deterioration of their health from light skin disease to severe skin disease and from no skin disease to severe skin disease. Theoretically, the willingness to pay and the willingness to accept questions should produce (approximately) identical answers.[1] Unfortunately, experimental research typically reports a wide divergence between the two measures. The WTP-WTA disparity is only one of the biases that distorts the elicitation of monetary values for health states [for an extensive overview see Hausman, 1993].

The difficulties encountered in the elicitation of monetary values for health benefits have stimulated the development of other types of analysis. The type of analysis which is now most common in health care decision making is cost utility analysis. In cost utility analysis the benefits of health care programs are not expressed in money terms, but in utility terms. The utility index that has been used most frequently is quality adjusted life years, written QALYs. The idea underlying QALYs is that life years should not enter unweighted in the calculation of health care benefits, but should be adjusted for the quality of life in which they are spent. To compute the number of QALYs, quality weights or utilities have to be assigned to health states. In the choice of treatment for skin disease, to perform a utility based analysis utilities have to be assigned to the three relevant health states: no skin disease, light skin disease and severe skin disease. Suppose that somehow we have managed to determine that the individual's utility of severe skin disease (denoted in the sequel by *u(severe skin disease)* is equal to 0.50. Further *u(light skin disease)* = 0.80 and *u(no skin disease)* = 1. Then the number of QALYs associated with treatment 1 is *10 years*0.8=8*. The number of QALYs associated with treatment 2 is *2*0.50+8*1=9*. If an individual has to make a choice between the two treatments and he is not concerned about costs, the QALY criterion dictates that he should choose the second treatment.

The above argument can be expressed in a more formal way. If the implementation of a particular health care program results in T life years spent in varying levels of quality of life q_t, where q_t denotes quality of life in period t, then the number of QALYs generated by this program is calculated as

---

[1] Some deviation is allowed because of substitution effects [Hanemann, 1991].

$$\sum_{t=1}^{T} u(q_t) \qquad (1)$$

where the function $u(q_t)$, which can be interpreted as a utility function over quality of life, assigns quality weights, or utilities, to health states. In the above example we have $u(severe\ skin\ disease) = 0.5$ and $u(light\ skin\ disease) = 0.8$.

No consensus exists as to the appropriate way to elicit the quality weights $u(q_t)$ in equation (1). Three principal methods have been used: the rating scale, the time trade-off and the standard gamble [Drummond, Stoddart and Torrance, 1987; or chapter 6 of this thesis]. Unfortunately, empirical evidence has shown that these three methods lead to different utilities [e.g. Torrance, 1976; Read et al., 1984; Hornberger et al., 1992; chapter 6]. This disparity between utilities may have the unfortunate consequence that choices between treatments vary depending on the method used to elicit the utilities. For example, if in the choice problem of treatment for skin disease described above $u(severe\ skin\ disease) = 0.45$ and $u(light\ skin\ disease) = 0.90$, the reader can easily verify that treatment 1 will be preferred. The fact that different methods lead to different utilities underlines the need to determine which of the three methods should be preferred. In the early years of cost utility analysis the view was widely shared that the standard gamble is the preferred method [Torrance, 1976; Weinstein, Fineberg et al., 1980]. The argument underlying the alleged superiority of the standard gamble was that the standard gamble is based on an axiomatic theory of decision under risk which has a wide appeal both as a description of individual preferences (descriptive validity) and as a theory of how rational individuals should ideally behave (normative validity): von Neumann and Morgenstern's expected utility theory (1944). However, over the past decades the appeal of expected utility theory has been challenged. A stream of papers starting with Allais (1953) has shown that expected utility theory may fail as a descriptive theory of decision under risk. Moreover, health research has indicated that the standard gamble leads to theoretically inconsistent results in health state utility measurement [cf. e.g. Llewellyn-Thomas et al., 1982; Rutten-van Mölken et al., 1995]. As a result of these studies an increasing number of researchers have lost their confidence in the standard gamble as the "gold standard" in health state utility measurement. To date, none of the other two methods has replaced the standard gamble as the "gold standard." This has left health state utility measurement in a state of disarray: several methods exist for the elicitation of health state utilities, evidence abounds that, given common scaling, these methods lead to significantly different utilities, but little is known about the relative merits of the methods.

Even though equation (1) reflects the basic idea behind QALYs (life years should not enter unweighted in the evaluation of the benefits of health care

programs, but should be adjusted for the quality in which they are spent), it is not the model that is most frequently used in practical applications. In the choice of treatment for skin disease the health benefits are realized at different points in time. Most analyses adjust costs and benefits for such differences in time of realization. The rationale behind this adjustment is that the timing of outcomes affects the attractiveness of the outcomes. For example, when asked to make a choice between 10 guilders now and 10 guilders in 1 year's time, most individuals will prefer 10 guilders now. Economists refer to this phenomenon as time preference. The general pattern is that the utility derived from an outcome is higher the sooner it occurs. This type of behaviour is defined as positive time preference, as opposed to negative time preference in which outcomes are more preferred the later in time they occur. To allow for positive time preference the weight that is attached to an outcome should decrease with time. Time preference is commonly accounted for by applying a constant rate of discount *(r)*. Constant rate discounting implies that equation (1) should be replaced by:

$$\sum_{t=1}^{T} \frac{u(q_t)}{(1+r)^{t-1}} \qquad (2)$$

Let us return to the choice of treatment for skin disease. Suppose we decide to apply a discount rate of 5% per year, that is, r in equation (2) is equal to 0.05. Then the number of discounted QALYs associated with treatment 1 is 6.49 and the number of discounted QALYs associated with treatment 2 is 7.13. Treatment 2 is still preferred by the criterion "maximize the number of 5% discounted QALYs," but the difference between the two treatments has become smaller. This could be expected given that treatment 2 gains its comparative advantage over treatment 1 only from period 3 onwards.

Thus far, I have spoken only about individual decision making. However, a second important aim of cost utility analysis is to guide societal decision making with respect to the allocation of health care resources. Choices have to be made between programs affecting several individuals. To make such choices, a procedure has to be determined to aggregate benefits over individuals. The common approach to the determination of the societal benefits generated by a health care program, is by unweighted summation of QALYs over all individuals affected by the health care program. Denote the number of individuals affected by a program by I. The total number of QALYs generated by this program is then calculated as:

$$\sum_{i=1}^{I} \sum_{t_i=1}^{T_i} \frac{u(q_{t_i})}{(1+r)^{t_i-1}} \qquad (3)$$

where the subscript i in t_i reflects the fact that the number of life years generated by the health care program need not necessarily be equal across individuals. Strictly speaking u should also be individual-specific. However, in practical applications the commonly employed procedure is to use the mean value of u and to hypothesize that this applies to all individuals involved.

Suppose, in the example of treatment for skin disease, that treatment 1 is cheaper than treatment 2. More specifically, for every 1000 patients that are treated by treatment 2, 1075 patients can be treated by treatment 1. Suppose that all patients have a life-expectancy of at least 10 more years and ignore the fact that after these 10 years some patients may receive additional treatment. Finally, suppose that the mean utility weights are similar to the individual utility weights given above. Then, applying equation (3), the number of QALYs per given amount of costs generated by a program that offers treatment 1 to 1075 patients is 6976.75. The number of QALYs generated by a program that offers treatment 2 to 1000 patients is 7130. The second program is to be preferred on the basis of the criterion "maximize the number of (discounted) QALYs for a given amount of costs."

1.2 Questions

Equations (1), (2), and (3) summarize QALY based decision making. In individual treatment decisions, the QALY model recommends the treatment that maximizes equation (1) or equation (2) (depending on whether or not time preference should be incorporated in the analysis). In societal decisions with respect to the allocation of resources over health care programs, the QALY model dictates that the program should be chosen that maximizes equation (3) for a given amount of costs.

Since its introduction in the seventies, the number of practical applications of QALY based decision making increased rapidly in the 1980s. This increase in practical applications was not matched by research into the theoretical properties of QALYs. Even though attention was drawn at various places in the literature to remaining problems associated with the use of QALYs [e.g. Torrance, 1986], the concept itself remained unchallenged. However, in 1989 two papers were published that strongly criticized QALY based decision making [Mehrez and Gafni, 1989; Loomes and McKenzie, 1989]. These two papers argued that QALY based decision making, both at the individual and at the societal level, relies on restrictive

assumptions that will generally not be fulfilled. Consider again the skin disease example. Using equation (2) I predicted that an individual should choose treatment 2. However, this conclusion depends crucially on whether the individual's preferences can indeed be described by equation (2). It might well be that if the individual is asked to make a direct choice between the two treatments, he would select treatment 1. In this case, the QALY model does not provide a correct description of the individual's preferences.

The criticism of Mehrez and Gafni and of Loomes and McKenzie led among health care researchers both to a greater awareness of the limitations of QALY based decision making and to confusion as to the fundamentals of QALY based decision making. This mood of confusion is well reflected by the title "QALYs: where next?" of a paper by Mooney and Olsen (1991). These authors identify various areas that need clarification and elaboration before one can have sufficient confidence in QALY based decision making both at the individual and at the societal level.

The state of confusion in the literature has led to the work on this thesis. The aim is to provide health utility indices in general and QALYs in particular with a foundation in the economics of decision theory. The basic data that decision theory seeks to explain are preference relations. To make preference relations tractable, conditions are imposed that allow these preference relations to be described by representing functions. In the skin disease example I used the models in equations (1) and (2) as a representing function to describe the choices of an individual between two different treatment options. These equations can only be representing functions if the individual's preferences for health satisfy certain conditions. One of the aims of this thesis is to identify these conditions. The advantage of identifying preference conditions is that these conditions allow an empirical assessment of the representing function. By means of the conditions it is possible to examine to what extent a representing function is normatively and descriptively valid. Normative validity refers to the question whether it is rational for a decision maker to behave according to the representing function. Descriptive validity refers to the question whether actual behaviour satisfies the representing function. To clarify the distinction between normative and descriptive validity consider the example of the treatment for skin disease. Suppose we have identified the conditions under which an individual's preferences can be represented by equation (2). One such condition can for example be: if the individual prefers health profile A (e.g. 10 years without skin disease) to health profile B (e.g. 10 years with light skin disease) and he also prefers health profile B to health profile C (e.g. 10 years with severe skin disease) then the individual should also prefer health profile A to health profile C. This preference condition is referred to as transitivity. Suppose that after examining this

condition we conclude that it is reasonable to expect the individual to behave according to transitivity. Then we conclude that transitivity is normatively appealing and a model that only depends on transitivity is normatively valid. However, that does not mean that the individual will behave according to transitivity. It may be that in actual choice situations the individual violates transitivity systematically. In that case we conclude that transitivity is not descriptively valid: it does not provide a good description of the individual's preferences in this choice context.

This thesis interprets the QALY as a utility model, which represents a preference relation. Throughout chapters 2 to 8 the underlying preference relation is the individual preference relation. Obviously, the results derived in these chapter have relevance for the use of QALYs in individual medical decision making. In this context, the aim is to assist an individual patient to choose the treatment option that is in accordance with his preferences. It is obvious that any utility model that is used in the context of individual decision making should ideally be consistent with individual preferences.

In the context of societal decisions with respect to health care programs, the relevance of the results derived in chapters 2 to 8 is more subtle. The focus is now on social preference relations and it is not immediately clear what influence the preferences of the individuals that constitute society have on these social preference relations. Broadly speaking two interpretations of QALYs as societal decision rules can be distinguished in the literature: QALYs as measures of health and QALYs as utilities. Even though these two interpretations are not necessarily mutually exclusive in the sense that QALYs as utilities can also be measures of health, the difference between the two interpretations rests on the question whether or not QALYs reflect underlying individual preference relations. In the former interpretation QALYs are defined without reference to individual preferences. In the second interpretation individual preferences determine health state utilities and these are subsequently aggregated into social utilities. This latter interpretation, which is supported by Torrance (1986) among others, is in line with the literature on social choice theory, in which social welfare functions are constructed from individual preferences. Clearly for the interpretation of QALYs as utilities, the study of individual preference relations is important. This thesis is therefore in line with the second interpretation of QALYs: QALYs as social decision rules should be a reflection of individual preferences. In fact, I feel that this interpretation is implicitly supported by most researchers in the field. Otherwise, it is hard to understand why such a large body of research is devoted to finding appropriate techniques to elicit individual preferences for health states.

Even though the results of this thesis mainly have relevance for the interpretation of QALYs as utilities, the analysis in chapter 9 is also applicable to the interpretation of QALYs as measures of health.

Given the focus of this thesis on QALYs as functions that represent preferences for health, four central questions can be formulated:

1. Under what conditions is the QALY model expressed in equation (1) a valid representation of the individual preference relation over sequences of health outcomes? Are these conditions descriptively and normatively appealing? If not, do alternative models perform better?
2. If the objective is to describe individual preference relations, which of the three methods (rating scale, time trade-off and standard gamble) elicits quality weights that describe individual preferences best?
3. If individual preferences depend on the temporal realization of outcomes, under what conditions can the intertemporal preference relation be described by the constant rate discounted utility model represented by equation (2)? Are these conditions descriptively and normatively appealing? If not, do alternative ways of modelling intertemporal preferences explain intertemporal behaviour better?
4. What conditions have to be imposed on the social preference relation to ensure that social preferences can be represented by equation (3), i.e. unweighted aggregation? Do these conditions provide an appropriate description of social preferences? If not, do alternative social decision rules perform better?

1.3 Structure

1.3.1 Validity of QALYs

Chapters 2 to 4 address the first of these four central questions. In chapter 2 I consider the simplest case where sequences of health outcomes (health profiles) are of constant quality, i.e. all q_t are equal. An example of a health profile of constant quality is treatment option 1 in the choice of treatment for skin disease described above. In the first option the resulting health profile consists of 10 years with light skin disease, i.e. all q_t are equal to light skin disease. In an influential paper, Pliskin, Shepard and Weinstein (1980) have derived under what conditions the individual preference relation over (lotteries over) health profiles can be represented by equation (1). In chapter 2 the analysis by Pliskin, Shepard and Weinstein is generalized. It is shown that equation (1) can be derived by imposing only one of

their conditions, risk neutrality on life years, if the preference relation simultaneously satisfies a condition that is entirely plausible in the medical context. Chapter 3 extends the analysis of chapter 2 to the case where health profiles are allowed to be of a varying quality, i.e. not all q_t are necessarily equal. An example of a health profile of varying quality is treatment option 2 in the choice of treatment for skin disease. In treatment option 2 the first two years are spent with severe skin disease, the next 8 without skin disease. The preference conditions derived in chapter 2 are not sufficient to represent choices for health profiles of varying quality by the QALY model. I show in chapter 3 what conditions have to be imposed on the preference relation to derive the QALY model in the case of health profiles of varying quality. Concern about the validity of the conditions underlying the QALY model has prompted Mehrez and Gafni (1989) to propose an alternative outcome measure: the healthy-years equivalents (HYEs). Chapter 3 also contains an assessment of this alternative outcome measure. It is argued that even though HYEs accommodate a more general class of preferences, their practical applicability is problematic. Chapter 3 concludes by proposing a utility index that is theoretically less restrictive than QALYs and easier to apply in practice than HYEs.

Chapter 4 describes the results of an experiment designed to examine to what extent respondents' preferences satisfy two conditions identified by Pliskin, Shepard and Weinstein. These two preference conditions underly a more general utility measure. Insight in the extent to which these two preference conditions hold, provides insight in the descriptive validity of this more general utility measure.

1.3.2 Which U?

Chapters 5 and 6 address the second question formulated above, which of the three methods to determine health state utilities should be preferred. Chapter 5 examines an inconsistency in standard gamble valuations: theoretically equivalent gamble questions elicit utilities that are not only different, but that differ in a systematic manner. Three phenomena are considered that may explain this systematic disparity. The first explanation focuses on the evaluation processes that may differ among the gamble questions (framing bias). Psychological research has shown that asking different questions induces different cognitive processes. These cognitive processes may explain the observed disparity. The second explanation attributes the disparity to imprecision in the preferences of respondents. In standard gamble questions respondents are confronted with a task and with health states they are not familiar with, which is likely to cause a certain degree of imprecision in their preferences. This imprecision may result in the observation of the disparity. The

hypothesis tested in chapter 5 is that the disparity is an artifact of the imprecision of preferences and will disappear once imprecision of preferences is taken into account. The third explanation relates to a recently proposed theory of decision under risk: rank dependent utility theory. Rank dependent utility theory distinguishes itself from expected utility theory in that probabilities are transformed into decision weights. It is well known from psychological research that individuals do not enter probabilities linearly in their evaluation of gambles. Rather they have a tendency to overweight small probabilities, and to underweight probabilities in the middle range. This type of behaviour explains for example why individuals who are generally averse to risk participate in lotteries giving a small probability of success.

Experimental data are presented that provide insight in the contribution that the three explanations can make to the explanation of the observed disparities between theoretically equivalent gambles. Given the observed inconsistency of the standard gamble, the argument that the standard gamble can be considered to be the norm in health state utility measurement against which the performance of rating scale and time trade-off should be assessed, can be questioned. There are no a priori theoretical reasons to prefer time trade-off or rating scale either. Given its embedding in decision theory, this thesis proposes to take the individual preference relation as the standard against which the performance of the three methods should be measured. The central question then becomes which of the three QALY models (QALYs based on rating scale weights, QALYs based on time trade-off weights, or QALYs based on standard gamble weights) is most consistent with individual preferences. An experiment was designed to examine this question. Chapter 6 reports the results of this experiment.

1.3.3 Time preference

Intertemporal decision making is the subject matter of chapters 7 and 8. Those two chapters discuss the third question, whether individual intertemporal preferences for health should be represented by the constant rate discounted utility model and if not, if alternative theories exist that are better able to explain intertemporal preferences for health. Chapter 7, which is theoretical, provides a characterization of the constant rate discounted utility model and assesses the descriptive and normative validity of the underlying preference conditions. Chapter 8 is empirical. It describes the results of an experimental test of the central condition of the constant rate discounted utility model. Chapter 8 further presents a brief description of a class of alternative intertemporal models to which I refer as generalized discounted utility models. The difference between constant rate

discounted utility models and generalized discounted utility models is that the latter attach more weight to future outcomes. Chapter 8 describes the key property on which these generalized discounted utility models are based. I show in chapter 8 what this key property implies in terms of the experiment described there. This allows to draw inferences with respect to the descriptive validity of generalized discounted utility models.

1.3.4 Equity

Chapter 9 differs from the preceding chapters in that it takes the social preference relation rather than the individual preference relation as primitive. Chapter 9 discusses the fourth question described above, whether unweighted aggregation, also referred to as "QALY-utilitarianism", is a correct representation of social preferences with respect to QALY (or health utilility indices in general) allocations. I emphasized before that this thesis interprets QALYs as a utility measure based on the aggregation of individual preferences. The interpersonal comparability of individual utilities has been a topic of fierce debate in economics. Influential authors such as Arrow have argued against the possibility of aggregating individual utilities. In chapter 9 a rationale is presented why such aggregation may be applied.

The possibility of aggregating QALYs over individuals does not imply that this aggregation should take the form of unweighted aggregation. In the skin disease example, unweighted aggregation resulted in a social preference for treatment 2. However, unweighted aggregation ignores the fact that the program pursuing treatment 1 resulted in 1075 patients being treated for a given amount of costs, whereas the program pursuing treatment 2 only resulted in 1000 patients being treated. That is, in the latter program 75 patients are left untreated, i.e. they have to live 10 years with severe skin disease. Out of equity considerations a policy maker may consider this situation undesirable. If a policy maker is concerned about equity, this concern should be reflected in the social decision function. Unweighted aggregation cannot capture such concerns. It reflects the idea that "a QALY is a QALY no matter who gets it." Chapter 9 proposes alternative utility indices that incorporate equity concerns.

Chapter 10 contains a discussion of the conclusions derived in this thesis and the direction future research in this area may take.

The chapters in this thesis have been written as separate articles for journals. This has the advantage that each chapter can be read independently of other chapters and

that the reader can skip chapters that he is not interested in. A disadvantage is that there may be a certain overlap between the chapters.

The ordering of the chapters is the one which seemed most logical given the four questions this thesis is based on. The chapters are therefore not ordered chronologically in the sense that chapters with a lower number were written first. By consequence, results that are derived in earlier chapters were not necessarily known when writing later chapters. This should be kept in mind when reading this thesis.

2

Characterizing QALYs by means of risk neutrality[1]

Summary

This chapter shows that QALYs can be derived from more elementary conditions than thought hitherto in the literature: it suffices to impose risk neutrality for life years in every health state. This derivation of QALYs is appealing because it does not require knowledge of concepts from utility theory such as utility independence -risk neutrality is a well known condition. Therefore our axiomatization greatly facilitates the assessment of the normative validity of QALYs in medical decision making. Moreover, risk neutrality can easily be tested in experimental designs, which makes it straightforward to assess the descriptive (non)validity of QALYs.

2.1 Introduction

Quality-adjusted life years (QALYs) are the most common outcome measure in cost utility analysis. They offer a straightforward procedure for combining the two most important outcomes of health care programs, quality of life and quantity of life, into one single measure. QALYs have the advantages of being easy to calculate, and having an intuitively appealing interpretation. A disadvantage is that they require the individual preference relation to satisfy some restrictive conditions. Given the importance of QALYs and the many discussions of their appropriateness, further insights into those restrictive conditions is important.

The aim of this chapter is to provide a characterization of QALYs for the case of chronic health states that is more elementary and fundamental than those provided hitherto in the literature. Throughout we assume expected utility. The

[1] Based on Bleichrodt, H., P. Wakker, and M. Johannesson: "Characterizing QALYs by means of risk neutrality" (submitted for publication).

conditions commonly used to characterize QALYs are "utility independence," "constant proportional tradeoffs," and "risk neutrality for life years." These conditions were established by Pliskin, Shepard, and Weinstein (1980), and studied also by others [cf. e.g. Torrance and Feeny, 1989; Loomes and McKenzie, 1989; Mehrez and Gafni, 1989; Culyer and Wagstaff, 1993; Bleichrodt, 1995]. The surprising result provided here is that, in the presence of a condition that is unobjectionable in the medical context, the condition of risk neutrality for all health states alone already suffices to imply QALYs. That is, in the medical context risk neutrality simply implies the other two conditions.

Representation theorems aim to identify the conditions that underly a particular preference representation. This is important both for normative and for descriptive reasons. Normatively, by examining the preference conditions a decision maker can be persuaded to use a particular model or, alternatively, the preference conditions can be used as an argument for not using a model. Descriptively, identifying the preference conditions allows for the testing of the model in an experimental setting. The attractiveness of a particular representation depends crucially on the conditions used. Conditions that are easy to understand and/or intuitively appealing facilitate the tasks of assessing the normative and descriptive properties of a model.

The central condition in our characterization, risk neutrality with respect to life years, is well-known and can easily be explained. It does not require knowledge of utility theory concepts such as utility independence. Thus our result is both more elementary and more general than the existing results in the literature. Also, by finding a shorter road to QALYs, we can provide an extremely simple proof that is easily illustrated graphically. The proof is so simple that it is given in the main text. We hope that all readers, not only those acquainted with utility theory, will be able to understand the characterization of the QALY model, and also the proof thereof.

After the presentation of the main theorem of this chapter, we provide a detailed analysis of the relations between our conditions and the ones customary in the literature. This will further clarify the points where we generalize existing results.

2.2 Structural assumptions

We restrict attention to chronic health states in this note, that is, we assume that quality of life is constant until death. Thus a pair (Q, T) designates the outcome where a person lives for T years in health state Q and then dies. We adopt the structural assumptions commonly used in the study of multiattribute utility and

medical decision making [Keeney and Raiffa, 1976; Pliskin et al., 1980]. That is, we study an individual preference relation on lotteries over chronic health states. By $[p_1,(Q_1,T_1);p_2,(Q_2,T_2);\ldots;p_n,(Q_n,T_n)]$ we denote a lottery yielding outcome (Q_i,T_i) with probability p_i. The preference relation satisfies the von Neumann-Morgenstern axioms [von Neumann and Morgenstern, 1944]. Hence there exists a utility function U, assigning to each chronic health state (Q,T) the utility $U(Q,T)$, such that the expectation of U, $(p_1*U(Q_1,T_1)+\ldots+p_n*U(Q_n,T_n)$ for the above lottery), governs the choices between lotteries over chronic health states.

2.3 The main result

Definition 2.1. The individual preference relation satisfies risk neutrality for life years if, with quality of life held fixed, the individual is indifferent between a lottery over life years and the expected life duration of that lottery.

Risk neutrality means that, for any particular health state Q, the individual is indifferent between:

(i) a probability p of T years in Q and a probability $(1-p)$ of S years in Q;
(ii) $p*T+(1-p)*S$ years in Q for certain.

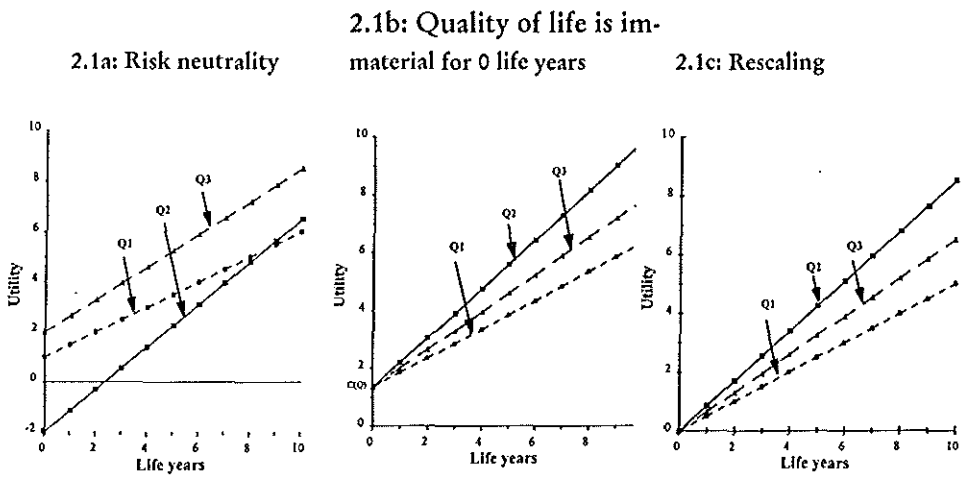
If we draw $U(Q,T)$, holding quality of life Q constant, then risk neutrality implies that the graph of $U(Q,T)$ is linear. Risk neutrality is illustrated in figure 2.1a where the utility function over life years has been drawn with quality of life held fixed at three different levels. Let us emphasize that linearity of utility is necessary and sufficient for risk neutrality, and does not require that the slope of utility be one or that the intercept be zero. Formally, by risk neutrality the von Neumann Morgenstern utility function $U(Q,T)$ is of the form $C(Q)+V(Q)*T$, where $C(Q)$ is a constant that depends on Q, but is independent of T and $V(Q)$ is a positive constant that depends on Q, but is independent of T. In figure 2.1a, $C(Q)$ is the intercept of $U(Q,T)$ and $V(Q)$ the slope.

It is obvious that under the QALY model, where $U(Q,T)=V(Q)*T$ for a function V, $U(Q,0)$ must be the same for all health states. That condition is not implied by the assumptions made so far. In particular, it is not satisfied in figure 2.1a. Therefore the QALY model fails in figure 2.1a. Let us display this condition.

Condition 0. For a time duration of zero life years all quality of life levels are equivalent.

Condition 0 is entirely self-evident in the medical context. Figure 2.1b illustrates the effect of imposing condition 0 in addition to risk neutrality on life years. By Condition 0, $U(Q,0)$ is constant for all health states. Thus $U(Q,0) = C(Q) + V(Q)*0 = C(Q)$ is constant. For a time duration of zero life years Condition 0 implies that al linear utility lines pass through the same point $C(Q)$, i.e. they must have the same intercept.

Figure 2.1: Illustration of a complete derivation of QALYs (the figure shows U(Q,T) for three different levels of quality of life)



It is well-known in the von Neumann-Morgenstern utility theory that one can add up, at one's will, a constant to a von Neumann-Morgenstern utility function. Therefore we can add up minus the intercept $C(Q)$ in figure 2.1b. That is, we may assume figure 2.1c in which $U(Q,0) = 0$ for all Q. We can now write

$$U(Q,T) = V(Q)*T \qquad (1)$$

The above equation has established the QALY model. Let us summarize: If expected utility, risk neutrality for each fixed health state, and Condition 0 hold, then the QALY model holds. It is obvious that the conditions are also necessary for the QALY model. Therefore we have established:

Theorem 2.1. Under expected utility, the following two statements are equivalent for a preference relation over chronic health states:

(i) The QALY model holds:
$U(Q,T) = V(Q)*T$.

(ii) Condition 0 holds and, for each health state, risk neutrality holds for life years.

Since Condition 0 is unobjectionable in the medical context, the above theorem has demonstrated that risk neutrality for all health states is the essence of the QALY model. The QALY model can be justified normatively if and only if risk neutrality can be, and it can be criticized normatively if and only if risk neutrality can be. Similarly, the QALY model can be verified descriptively if and only if risk neutrality can be, and it can be falsified if and only if risk neutrality can be. The general finding, both normatively and descriptively, is that risk neutrality does not hold [cf. e.g. McNeil et al., 1978; Maas and Wakker, 1994; Stiggelbout et al., 1994; Verhoef et al., 1994]. Therefore QALYs can at best be used as an approximation, in contexts where the violations of risk neutrality are not unacceptably extreme. The realism of the QALY model can be increased if the numbers T do not designate life years, but discounted life years [Johannesson et al., 1994] or powers of life years [Pliskin et al., 1980; Miyamoto and Eraker, 1985; Miyamoto and Eraker, 1988; Stiggelbout et al., 1994]. Obviously, such a different interpretation of the number T in our theorem does not change its mathematical correctness. Thus risk neutrality with respect to discounted life years holds if and only if the QALY model holds with discounted instead of absolute life years.

2.4 A comparison with the result of Pliskin et al.

The characterization of QALYs that is commonly invoked in the literature has been established by Pliskin et al. (1980). Instead of Condition 0, Pliskin et al. impose utility independence and constant proportional tradeoffs. One reason for Pliskin et al. to consider these conditions is that they can also serve to characterize models that are more general than the QALY model studied in this chapter. The importance of risk neutrality for QALY characterizations was already suggested by Johannesson (1995). That paper, however, did not provide a complete characterization and derivation. Below we first discuss constant proportional tradeoffs.

Definition 2.2. Let Q_1 and Q_2 be two health states. They satisfy the constant proportional tradeoffs assumption if there exists a positive number q such that $U(Q_1, T) = U(Q_2, qT)$ for all life durations T.

In other words, constant proportional tradeoffs hold if the proportion of life years the individual is willing to give up for a given quality of life improvement is invariant with respect to the remaining number of life years. Pliskin et al. imposed constant proportional tradeoffs only for a best and worst state of health, and then proved that in the presence of utility independence that implies constant proportional tradeoffs for all states of health. That, in turn, immediately implies our Condition 0, simply by substituting $T = 0$ in the above definition. This implication also demonstrates how the QALY axiomatization of Pliskin et al. can be derived from ours: One derives Condition 0 as just indicated, and then by risk neutrality and Theorem 2.1 the QALY model follows. Condition 0 can be viewed as a weakened version of constant proportional tradeoffs.

Let us, for completeness, point out that risk neutrality in isolation does not imply constant proportional tradeoffs: In Figure 2.1a, q is close to 1 for large T but remote from 1 for small T. Risk neutrality and Condition 0, however, do imply constant proportional tradeoffs as follows immediately from the representation $U = V(Q) * T$ in Theorem 2.1 (define $q = V(Q_1)/V(Q_2)$ in the above definition).

Next we discuss utility independence.

Definition 2.3.

- Quality of life is utility independent from quantity of life if preferences over lotteries for quality of life with quantity of life held fixed at level T are invariant with respect to the particular level T.

- Quantity of life is utility independent from quality of life if preferences over lotteries for quantity of life with quality of life held fixed at level Q are invariant with respect to the particular level Q.

- If both conditions hold, we say that quality of life and quantity of life are utility independent.

If quality of life is utility independent from quantity of life, then $[p,(Q_1, T); 1\text{-}p,(Q_2, T)]$ is preferred to (Q, T) if and only if, for any life duration T' different than T, $[p,(Q_1, T'); 1\text{-}p,(Q_2, T')]$ is preferred to (Q, T'). If quantity of life is utility independent from quality of life, then $[p,(Q, T_1); 1\text{-}p,(Q, T_2)]$ is preferred to (Q, T) if and only if, for any health state Q' different than Q, $[p,(Q', T_1); 1\text{-}p,(Q', T_2)]$ is preferred to (Q', T).

Obviously, if risk neutrality holds irrespectively of the quality of life, then for a fixed health state the preferences are governed by expected life duration, irrespective of the health state, and quantity of life is utility independent from quality of life. Conversely, if risk neutrality holds for perfect health and quantity of life is utility independent from quality of life, then risk neutrality holds for all qualities of life. This follows from the fact that by utility independence all utility functions over life years are strategically equivalent regardless at which level quality of life is held fixed. Thus, if risk neutrality holds for life years in full health, risk neutrality for life years must, by utility independence, hold for all health states. Therefore the following theorem is not surprising.

Theorem 2.2. Risk neutrality holds for all qualities of life if and only if quantity of life is utility independent from quality of life and risk neutrality holds for perfect health.

A remarkable implication of Theorem 2.1 is that risk neutrality, in the presence of Condition 0, does imply utility independence of quality of life from quantity of life. This is easily seen for the utility function $U(Q,T) = V(Q)*T$ in Theorem 2.1, because the expectation of $V(Q)$ governs preferences over qualities of life for a fixed level of T, independent of what that level of T is.[2]

Risk neutrality in isolation does not imply utility independence of quality of life from quantity of life. This can be seen as follows. Risk neutrality does not exclude $U(Q_1,5) > U(Q_2,5)$ and $U(Q_1,10) < U(Q_2,10)$; here $U(Q_1,.)$ has a larger intercept, but a smaller slope, than $U(Q_2,.)$, and the lines intersect between $T=5$ and $T=10$. The strict preferences $(Q_1,5) > (Q_2,5)$ and $(Q_1,10) < (Q_2,10)$ reveal that quality of life is not utility independent from quantity of life. This situation is graphically displayed in figure 2.1a.
We summarize the above discussion in the following corollary of Theorem 2.1 .

Theorem 2.3.
(i) Risk neutrality and Condition 0 imply both utility independence and constant proportional tradeoffs.
(ii) In the characterization of the QALY model by means of risk neutrality, utility independence, and constant proportional tradeoffs, the following generalizations are possible:

---

[2] A minor modification should be made that is implicitly assumed throughout this paper: Utility independence is restricted to the domain where the life duration 0 is excluded, and requires that all health states be positive. These points have sometimes been overlooked in the literature.

constant proportional tradeoffs can be weakened to Condition 0

and either

- utility independence can be dropped

or

- risk neutrality and utility independence can be weakened to risk neutrality for perfect health and utility independence of life years from health states.

Theorem 2.3 demonstrates that, for empirical investigations of the QALY model, tests of utility independence and constant proportional tradeoffs are tests of *implications* of risk neutrality.

2.5 Concluding remarks

In this chapter we have shown that QALYs can be derived from an individual preference relation that satisfies the von Neumann-Morgenstern axioms by imposing risk neutrality for life years and a very weak condition, that for a time duration of zero years all health states are equivalent (Condition 0). Given that Condition 0 is intuitively self-evident in the medical context, the crucial condition in our characterization is risk neutrality for life years. Risk neutrality for life years is a condition that is both easy to understand and straightforward to test in an experimental design. Empirical research generally indicates that risk neutrality on life years is violated to a certain degree.

It can be deduced from the analysis presented in this chapter that the widely held belief in medical decision making that utility independence, constant proportional tradeoffs and risk neutrality on life years all have to be imposed for characterizing the QALY model is overly restrictive. Theorem 2.3 shows that each of these conditions can be relaxed considerably. If Condition 0, self-evident in the medical context, is accepted, then Theorem 2.1 shows that two of the three common conditions, utility independence and constant proportional tradeoffs can simply be dropped.

The use of such an accessible condition as risk neutrality as central condition is the strength of our characterization. Our result facilitates the assessment of the normative and descriptive appeal of QALYs.

3

QALYs and HYEs: Under what conditions are they equivalent?[1]

Summary

This chapter examines what restrictions have to be imposed on the individual's preference structure for QALYs and HYEs to yield identical results. It is shown that using QALYs involves imposing three additional restrictions. Empirical evidence suggests that these restrictions cannot be expected to hold in all applications. The main problem in using HYEs appears to be practical. An alternative index is proposed, that may help to bridge the gap between QALYs and HYEs by combining to some extent the advantages of the two measures.

3.1 Introduction

The evaluation of health care programs involves both technical and value judgements. The value judgements concern mainly the trade-off between the two important outcomes of such programs: quality of life gained and quantity of life gained.
The QALY approach offers one way of incorporating these two benefits of health care programs into one single index measure: quality adjusted life years gained. On the basis of this index decisions concerning the allocation of resources in the health care sector can be made. The program that should be implemented is the one that offers the largest number of QALYs per dollar or, what is equivalent, the one that has lowest costs per QALY gained.

The QALY approach has been criticized by several authors (e.g. Loomes and McKenzie, 1989; Mehrez and Gafni, 1989). The essence of this criticism is that since QALYs rely on certain fairly restrictive assumptions (Pliskin, Shepard and Weinstein,

[1] Based on Bleichrodt, H., 1995, "QALYs and HYEs: Under what conditions are they equivalent?", Journal of Health Economics 14, 17-37.

1980; Weinstein, Fineberg et al., 1980) as a representation of individual preferences, care should be taken in using them in the evaluation of health care programs. Mehrez and Gafni (1989) propose an alternative index, the Healthy Years Equivalents (HYEs), which as they claim, fully represents patients' preferences, stemming from the way they are calculated from each individual's utility function. At the same time HYEs retain some attractive properties of QALYs: combining quality of life and quantity of life in a single index and being easy to interpret.

The aim of this chapter is to show how QALYs and HYEs are related to each other, that is, under what assumptions about the underlying preference structure they will give identical results. Both the certainty case and the uncertainty case will be considered. Moreover it will be argued that the claim that HYEs fully represent an individual's preferences is not completely true. Even HYEs make simplifying assumptions concerning the individual's preference structure. Besides this theoretical reservation, the main problem in implementing HYEs to evaluate health care programs appears to be practical. An alternative index is proposed, which may help to bridge the gap between QALYs and HYEs by combining to some extent the advantages of the two measures.

3.2 Quality-Adjusted Life Years

The basic QALY model, as it is typically encountered in the literature, is simple. Abstracting from discounting,[2] denote by q_t the health status level in period t, where it is assumed without loss of generality that each period lasts one year. Assume that health status levels form a continuum. The number of QALYs represented by the lifetime health stream $Q_T=(q_1,\ldots,q_t,\ldots,q_T)$ where T is the number of years to live from now on, is

$$QALY = \sum_{t=1}^{T} u(q_t) \tag{1}$$

[2]Two things should be noted. First, discounting can be accounted for within this framework by imposing an additional stationarity assumption (Koopmans, 1972). Second, it is not clear that discounting of QALYs is necessary. Following the suggestions made by Torrance and Feeny (1989), health state utilities are measured over the individual's lifetime. Measuring health state utilities this way clearly incorporates time preference.

where $u(q_t)$ is the utility associated with health status level q in period t.[3]
By computing (1) for the various programs and dividing this amount by the costs one arrives at the decision rule maximize QALYs per unit of costs.

In the rest of this chapter I will refer to equation (1) as the QALY model. I will derive under what conditions an individual's preferences can be represented by this utility function. One could object against this that by taking the basic version of the QALY model as the QALY model, it is a bit like assuming that the micro-economics of the firm is locked into the perfect competition assumptions. Meeting this objection requires an assessment of which assumptions are essential for the QALY concept and which are made for convenience (ease of measurement). I will briefly comment on this in section 7.

3.3 Healthy Years Equivalents

The calculation of HYEs is slightly more involved. Denote the lifetime health stream again by $Q_T=(q_1,\ldots,q_t,\ldots,q_T)$. The utility function over this stream is $u(Q_T)$, which represents the utility as viewed now by the individual. Denote q^* as the best health state (generally perfect health) and q^o as the worst health state (generally death[4]). Let H be the number of years in q^* and H^* be the healthy-years equivalent of Q_T. The problem is now to find H^* such that

$$u(Q_T)=u(Q_{H^*}) \tag{2}$$

where Q_{H^*} is a lifetime health profile with $q_t = q^*$ for $t=1,\ldots,H^*$ and $q_t = q^o$ for $t=H^*+1,\ldots, T$.

In case the von Neumann Morgenstern axioms hold, Mehrez and Gafni (1991) have shown that HYEs can be measured by a two-stage lottery-based procedure.

---

[3]It has been pointed out to me by one of the referees that this equation is not correct, given that in health care evaluation it is the gain in QALYs that has to be measured rather than the total number of QALYs. However, this does not affect the results of this paper. Differences between utilities are only meaningful if additional axioms on top of the conditions derived in this paper are imposed on the individual's preference structure (see for example Krantz et al., 1971).

[4]Death need not necessarily be the worst health state. However, health states worse than death cause major theoretical problems. In the context of the QALY model health states worse than death cause a violation of one of the assumptions underlying the QALY model: mutual preferential/utility independence (see below). In the HYE model such health states may lead to negative values for the number of HYEs associated with a health care program.

3.4 The certainty case: value functions

What conditions have to be imposed on the individual's preference structure to make QALYs and HYEs equivalent? First consider the case where the outcomes of a medical intervention are certain. Denote by $\succeq$ the preference relation "at least as preferred as" and by $\succ$ the preference relaion "strictly preferred to". In the case of certainty, an individual's preferences can ideally be captured by a value function which has the following properties:

$$\text{if } x \succeq y \text{ then } v(x) \geq v(y) \tag{3a}$$

and

$$\text{if } x \succ y \text{ then } v(x) > v(y) \tag{3b}$$

where x and y are vectors of attributes from which the individual derives value, one of which is health. If the individual's behaviour satisfies certain axioms (Debreu, 1954, 1964) such a value function can be shown to exist.

The QALY approach, again abstracting from discounting, assumes the following value function to measure individual preferences for a lifetime health stream Q_T:

$$v(Q_T) = \sum_{t=1}^{T} v(q_t) \tag{4}$$

which is an additively separable value function.

Before deriving the assumptions sufficient for such an additive form to be a correct representation of the individual's preferences under certainty, it is useful to introduce some terminology. Suppose that there are n attributes from which an individual derives value: $X_1,\ldots,X_n$. These attributes map each act a into a point $X(a)=[X_1(a),\ldots,X_n(a)]$ in the n-dimensional consequence space. Suppose further that the vector of attributes x can be subdivided into two subvectors y and z where

$$y=(x_1,\ldots,x_r) \quad \text{and} \quad z=(x_{r+1},\ldots,x_n) \tag{5}$$

Consider two values for the vector y, $y^1=(y_1^1,\ldots,y_r^1)$ and $y^2=(y_1^2,\ldots,y_r^2)$, and one for the vector z, $z^1=(z_{r+1}^1,\ldots,z_n^1)$ then

Definition 3.1: y^1 *is conditionally preferred or indifferent to* y^2 *given* z^1 *if and only if*

$$(y^1, z^1) \succeq (y^2, z^1) \tag{6}$$

Definition 3.2: the set of attributes $Y = \{X_1, \ldots, X_r\}$ *is preferentially independent of the complementary set* $Z = \{X_{r+1}, \ldots, X_n\}$ *if and only if the conditional preference structure for* y *given* z^1 *does not depend on* z^1*. Or, Y is preferentially independent of Z iff for some* z^1

$$[(y^1, z^1) \succeq (y^2, z^1)] \Leftrightarrow [(y^1, z) \succeq (y^2, z)], \quad \forall\ y^1, y^2, z \tag{7}$$

Definition 3.3: the attributes $X_1, \ldots, X_n$ *are mutually preferentially independent if every subset Y of these attributes is preferentially independent of its complement.*

Theorem 3.1 (Debreu, 1960): Iff the attributes $X_1, \ldots, X_n$ *are mutually preferentially independent and* $n \geq 3$*, the value function is of the additive form*

$$v(x_1, \ldots, x_n) = \sum_{i=1}^{n} v_i(x_i) \tag{8}$$

where v_i is an additive value function over X_i.

Generally, v and each of the single-attribute value functions v_i are scaled from 0 to 1. Following this scaling convention the following form of the value function results

$$v(x) = \sum_{i=1}^{n} \tau_i v_i(x_i) \tag{9}$$

where the τ_i are scaling constants.

In the case of QALYs, the attributes of the value function are health status levels in different years. Thus, the value function, which is now a value function for health, consists of T attributes, health status levels in the various years $(q_1, \ldots, q_T)$ over which single-period value functions v_t are defined.

To arrive at the basic QALY formulation two more assumptions besides mutual preferential independence have to be imposed on the individual's value function:[5]

1. stable preferences over lifetime, i.e. all single-period value functions are identical.
2. all scaling constants are equal; this implies that improvements in health are equally important across periods.

The number of HYEs in the certainty case is calculated by finding the value of H^* for which

$$v(Q_T) = v(Q_{H^*}) \tag{10}$$

where Q_T and Q_{H^*} are defined as before.

Since HYEs do not impose any additional[6] restrictions on the individual's value function for health, QALYs and HYEs will yield identical results under the three assumptions derived above.

Both the QALY approach and the HYE approach assume that the value function for health exists. That is, health and non-health attributes in the individual's overall value function are assumed to be mutually preferentially independent. In other words, preferences for lifetime health streams can be considered without consideration of other, non-health, attributes that bear value to the individual. Without this additional assumption, neither QALYs nor HYEs will correctly represent the individual's preferences under certainty.

3.5 The uncertainty case: utility functions

Assume that expected utility is the appropriate criterion to use in choosing among alternatives.[7] That is, an individual's preferences can be captured by a utility function

---

[5] Alternatively, the special case of formula (8) can be obtained by adding a symmetry preference condition: permuting the coordinates does not affect the indifference class. A similar condition is used in chapter 9.

[6] The HYE model, to give sensible results, has to impose a monotonicity condition: individual preferences have to be increasing with respect to healthy years of life. This assumption is implied by the QALY model.

[7] Recently, the use of expected utility as the appropriate decision criterion has been challenged. For an overview of some recent developments in modelling preferences under uncertainty see for example Karni and Schmeidler (1991).

which has the property that given two probability distributions A and B over the multi-attribute consequences, probability distribution A is at least as desirable as probability distribution B iff,

$$E_A\,[u(\tilde{x})] \geq E_B\,[u(\tilde{x})] \tag{11}$$

where E_A and E_B are expectation operators taken with respect to distributions A and B respectively and $\tilde{x}$ denotes a stochastic outcome vector.

The utility function which describes the individual's behaviour is unique up to a positive linear transformation. Note that in the degenerate case, where one of the consequences occurs with probability one, the utility function reduces to a value function. Thus, a utility function is by definition a value function, but a value function is not necessarily a utility function.

The QALY approach assumes that the utility of the lifetime health stream Q_T can be assessed by the following utility function:

$$u(Q_T) = \sum_{t=1}^{T} u(q_t) \tag{12}$$

which is an additively separable utility function.

An alternative way of representing the utility of the lifetime health stream Q_T is by considering the two-attribute utility function $u(T,q)$ where q is a constant health status level representing the stream of health status levels in the T years. q can be obtained, recall that health status has been assumed to form a continuum, by solving the following equation:

$$u(q,\ldots,q) = u(Q_T) \tag{13}$$

This is the approach followed by Pliskin, Shepard and Weinstein (1980) (PSW). PSW consider the QALY representation

$$u(T,Q) = T^{*}u(q) \tag{14}$$

and impose the following conditions on the individual's preference structure:

1. mutual utility independence between life years and health status (assuming the latter is constant across periods)
2. constant proportional trade-off of life years for health status
3. risk neutrality with respect to life years.

The definitions of utility independence and mutual utility independence are generalizations of definitions 3.2 and 3.3 to the case of preferences defined over lotteries over attributes rather than over the attributes themselves.

Definition 3.4: Attribute Y is utility independent of Z when conditional preference for lotteries on Y given Z do not depend on the particular level of Z.

Definition 3.5: Attributes Y and Z are mutually utility independent if Y is utility independent of Z and Z is utility independent of Y.

The assumption of a constant proportional trade-off of life years for health status implies that the proportion of remaining life years one is willing to give up for a given improvement in health status does not depend on the number of remaining life years. Risk neutrality over life years implies a linear utility function for life years.

Assume that the individual's utility function for health exists, that is health and non-health attributes in the individual's overall utility function are mutually utility independent. Then the three conditions are sufficient for the QALY model to be a correct representation of the individual's utility function for health.

Returning to equation (12), which is an alternative way of describing the utility function for a life time health stream, it can be asked under what conditions the intertemporal utility function reduces to this simple additive form. In answering this question, an alternative set of conditions is imposed on the utility function for health, which together have the same effect as the PSW conditions: a QALY can be considered a utility.[8] However, the analysis presented here is more general than the analysis of Pliskin, Shepard and Weinstein in that it is not confined to a derivation of the QALY model for chronic health states. Health profiles are allowed to vary over time.

Assume first that the attributes in the utility function for health (that is, health status levels in the various periods) be mutually utility independent. This assumption allows by applying theorem 6.1 in Keeney and Raiffa (1976) to write the utility function for health either in the additive form

$$u(q) = \sum_{t=1}^{T} k_t \, u_t(q_t) \quad \text{if} \quad \sum_{t=1}^{T} k_t = 1 \tag{15a}$$

[8] Again this is only true if health and non-health attributes in the individual's overall utility function are mutually utility independent.

or in the multiplicative form

$$1 + k u(q) = \prod_{t=1}^{T} [1 + k k_t \, u_t(q_t)] \quad \text{if} \quad \sum_{t=1}^{T} k_t \neq 1 \tag{15b}$$

where k and the various k_t are scaling constants and $u(q)$ and the various $u_t(q_t)$ are normalized such that $u(q^*) = u_t(q^*) = 1$ and $u(q^0) = u_t(q^0) = 0$.

In order to decide whether the additive or the multiplicative form is appropriate, the condition of additive independence has to be checked.

Definition 3.6: Attributes $q_1, \ldots, q_T$ are additive independent if preferences over lotteries on $q_1, \ldots, q_T$ depend only on their marginal probability distributions and not on their joint probability distribution.

Theorem 3.2 (Fishburn (1965, 1970): The additive utility function

$$u(q) = \sum_{t=1}^{T} k_t \, u_t(q_t) \tag{16}$$

is appropriate iff the attributes are additive independent.
where
i) u and u_t are normalized such that $u(q^*) = 1$; $u(q^0) = 0$; $u_t(q^*) = 1$ and $u_t(q^0) = 0$
ii) $k_t = u(q_1^0, \ldots, q_{t-1}^0, q_t^*, q_{t+1}^0, \ldots, q_T^0)$
A sketch of the proof has been given in the appendix.

Additive independence is a stronger restriction than mutual preferential independence. Additive independence implies mutual preferential independence, but mutual preferential independence does not imply additive independence. Denote a treatment scenario giving outcome (q_t^1, q_{t+1}^1) with probability p and outcome (q_t^2, q_{t+1}^2) with probability $(1-p)$ (suppressing periods in which health outcomes are equal) by $[(q_t^1, q_{t+1}^1), p, (q_t^2, q_{t+1}^2)]$. Imposing additive independence means that the individual is indifferent between the treatment scenarios: $T_1 = [(q_t^*, q_{t+1}^*), ½, (q_t^0, q_{t+1}^0)]$ and $T_2 = [(q_t^*, q_{t+1}^0), ½, (q_t^0, q_{t+1}^*)]$, for any two periods t and $t+1$ (again suppressing all other periods in which quality of life is assumed to be equal in the two outcomes of treatment). This is true, given that the marginal probabilities, which are the only basis for decision making in case additive independence holds, are equal and thus there should be indifference between the treatment scenarios.

The final two conditions that have to be imposed are, as in the case of preferences under certainty, stable preferences over lifetime and equal scaling constants. Stable preferences imply that the utility function for health can be written as

$$u(q) = \sum_{t=1}^{T} k_t \, u(q_t) \tag{17}$$

The assumption of equal scaling constants implies that

$$cu(q) = \sum_{t=1}^{T} u(q_t) \tag{18}$$

where $cu(q)$ measures the number of QALY's and $c = 1/k_1 = 1/k_2 = \ldots = 1/k_T$. Note that $cu(q)$ is still a utility function, since the utility function is unique up to a positive linear transformation and $c = 1/k_t = 1/(1/T) = T > 1$. As in the certainty case, equal scaling constants imply that improvements in health status are equally important across periods.

Since HYEs impose no additional[9] restrictions on the individual's utility function for health, the following assumptions are sufficient for QALYs and HYEs to yield identical results:

1. additive independence of health status levels in the various periods;
2. stable intertemporal preferences for health status;
3. equal weights attached to health improvements in various periods.

Both the QALY-approach and the HYE-approach assume that the single-attribute utility function for health exists. That is, health and non-health attributes in the individual's overall intertemporal utility function should be mutually utility independent.

[9]Again it has to be assumed that the utility function is increasing in healthy years.

3.6 Recent challenges of the HYE model

Recently the HYE approach has been challenged by various authors (Buckingham, 1993; Culyer and Wagstaff, 1993; Johannesson et al., 1993). Taken together, these authors raise three important points of criticism against the HYE:

1. The two-stage procedure to measure HYEs is nothing more than a complicated way of asking the time trade-off (Buckingham; Culyer and Wagstaff; Johannesson et al.)
2. HYEs are identical to QALY scores obtained from a time trade-off experiment (Culyer and Wagstaff).
3. HYEs assume risk neutrality with respect to healthy years (Johannesson et al.)

Since I have argued in sections 4 and 5 that QALYs and HYEs are only equivalent under certain restrictive assumptions, the above papers clearly challenge the results derived thus far.

The first point mentioned above, i.e. the assertion that measuring HYEs by the two stage standard gamble method proposed in Mehrez and Gafni (1991) will give identical result as measuring HYEs by a TTO question, is strictly speaking not the topic of this chapter, since the point is directed at the measurement procedure of the HYE rather than at the concept itself. Gafni et al. (1993) have responded to this objection by pointing out that measuring HYEs by the TTO method establishes equality between $v(Q_{H*})$ and $v(Q_T)$ whereas measuring HYEs by the two-stage procedure establishes equality between $u(Q_{H*})$ and $u(Q_T)$. It should be clear from sections 4 and 5 that in general value functions and utility functions are not identical and that there is no straightforward relationship between the two. Dyer and Sarin (1979, theorems 4 and 5) have proved that a (measurable) value function and a vNM utility function are only equivalent in the special case where mutual preferential independence and additive independence of the attributes are satisfied. However, from this argument it does not follow that HYEs elicited by the two stage procedure will differ from HYEs elicited by the TTO procedure. Value functions and utility functions differ in that they (generally) assign a different number to the same indifference class of consequences. This does not imply though that the consequences contained in an indifference class will be different. Indeed, as Loomes (1995) shows, under transitivity and monotonicity with respect to healthy years the number of HYEs elicited by the two stage procedure will be equal to the number of HYEs elicited by the TTO method.

From the equivalence of the two measurement methods it does not follow that the second point of criticism, raised by Culyer and Wagstaff, that HYEs are as

restrictive as QALY scores obtained from a time trade-off experiment, is correct though. Both the TTO-based QALY and the HYE start by reducing health scenarios to their equivalent number of years in perfect health by eliciting the value of H^* for which $v(Q_{H^*}) = v(Q_T)$. This determines the number of HYEs and as long as monotonicity with respect to healthy years holds, one can use this number to consistently rank health scenarios according to the individual's preferences. The TTO-based QALY approach makes two additional assumptions with respect to the individual's preference structure. First it is assumed that the profile Q_T can be reduced to a profile of constant health status (q^1), which is a continuity assumption with respect to health status, and second that the value H^*/T can be attached to q^1 and can be used in subsequent analyses where the number of years in q^1 is not necessarily equal to T. This is only justified if the assumptions outlined in section 4 hold.

Finally, the claim made by Johannesson et al., that HYEs assume risk neutrality with respect to healthy life years, is based on the conception that a HYE is a utility, that can be used in expected utility calculations, rather than an argument in the utility function. Even though it has been suggested in the literature that HYEs can be used in decision tree analyses and are thus to be interpreted as expected utilities [Gafni and Zylak, 1990], the HYE as proposed by Mehrez and Gafni (1989) is not intended to be a utility. If the utility function for health is increasing in healthy years, HYEs will be a correct representation of the individual preferences for health without having to impose any further restriction on this utility function. As outlined above, in this chapter we interpret HYEs to be an argument in the utility function and not a utility itself. Therefore the results of this chapter do not contradict the claim by Johannesson et al..

3.7 An assessment of the various assumptions

Having argued that HYEs and QALYs can theoretically differ, it can be asked whether it is likely that in practice they will differ. This amounts to an empirical assessment of the various assumptions. The present section will show that the QALY assumptions have been violated. It should be borne in mind though that the empirical evidence relates to the basic QALY model. This raises the question about what is intrinsic to a QALY. In my view, given that a QALY is essentially a weighting scheme in which life years are adjusted for quality of life levels, separability of life years and quality of life is essential to the QALY. This makes the independence assumptions essential for a QALY-type of health outcome measure. The assumptions of stable preferences and equal weights to health improvements are merely convenience assumptions which could be relaxed. Note though that these

"convenience assumptions" are typically made in empirical work, given that in their absence the assessment task becomes highly involved.

3.7.1 The certainty case

It has been shown in section 4 that under certainty, the QALY model relies on three assumptions: mutual prefential independence, stable intertemporal preferences and equal weights being attached to health improvements.

Mutual preferential independence over periods implies that preferences for health status in any two subsequent periods do not depend on the levels of health status in the other periods. Note first that this assumption is in clear conflict with one of the major contributions to the theory of health economics: Michael Grossman's (1972) model of the demand for health. Moreover, intuitively one would not expect mutual preferential independence to hold for every health profile. Especially in case preferences between the pairs (q_t^1, q_{t+1}^1) and (q_t^0, q_{t+1}^0) are being considered with q^1 denoting a health status level only slightly preferred to death, one would expect these preferences to be influenced by health status levels in other periods.

Evidence of this has been reported by Sutherland et al. (1982). Sutherland et al. showed that attitudes toward survival in various health states change with the time of additional increments in survival in these health states. A majority of their subjects preferred three months of survival (followed by death) to immediate death, even in highly dysfunctional health states. However, for the most dysfunctional health states a majority of subjects preferred immediate death to 8 years of survival in these health states (followed by death). These findings suggest the existence of some sort of threshold ("maximal endurable time" as Sutherland et al. describe this) above which increments in survival are negatively valued. Note that the existence of a "maximal endurable time" is in conflict with mutual preferential independence: preferences for health states within a period cannot be considered without taking into account health status levels in other periods. It also suggests that mutual preferential independence is most likely to hold either when the deviation from normal health takes place for a relatively short period or when the deviation is not severe. Unfortunately it is precisely in the evaluation of severe and/or chronic conditions that the QALY approach has typically been applied.

It is easy to show that equal scaling constants τ_i and stable preferences over time together imply a constant proportional trade-off of life years for health status. The available empirical evidence does not support this constant proportional trade-off assumption. Sackett and Torrance (1978), measuring the values of different health states by the time trade-off technique, found evidence of an increasing rather than a

constant proportional trade-off: the value both patients and members from the general public assigned to various health states decreased dramatically with increases in the number of years spent in those health states.

Pliskin et al. (1980) also confronted the subjects in their study with some time trade-off questions. Out of 30 questions only 9 answers were consistent with a constant proportional trade-off. Of these 9 consistent answers, 5 were such that the subject indicated that he/she was willing to trade-off no life years at all against improvements in health status. Eliminating these cases leaves only 4 out of 25 cases that were consistent with the constant proportional trade-off assumption.

If the constant proportional trade-off assumption is not tenable, it can be asked which of the two constituent assumptions, stable intertemporal preferences or equal weights being given to health improvements across periods, is most likely to cause the violation. Positive time preference has the effect of imposing different values for the τ_i's. More precisely, less weight will be given to health status levels further away in time and this can explain why increasing proportional trade-offs have been observed. The common practice of discounting QALYs can capture this phenomenon.

However, discounting is unlikely to solve all problems. As empirical evidence suggests (Loomes and McKenzie, 1989), people attach different weights to being healthy at various stages of the life-cycle. Ideally the scaling factors should reflect this by giving more weight to the value function for health status in those years (generally childhood and early parenthood). Simply discounting QALYs will not adequately represent this.

Concerning the stability of preferences over time, Sackett and Torrance (1978) found that elderly have somewhat different preferences for health status as assessed by the time trade-off method than younger people. The differences are small though. Moreover, even if significant differences would have been obtained, this would only have counted as weak evidence. Strong evidence can only be obtained by cohort studies. To date no such studies have been reported.

3.7.2 The uncertainty case

In section 5 it has been shown that under uncertainty the QALY model imposes additive independence, stable intertemporal preferences and equal weights for equal health improvements across periods on the individual's utility function.

Additive independence is a strong condition. It implies indifference between level health status streams and single-period health status. That is, the life-time certainty equivalent $q^+ = (q,\ldots,q)$ for the treatment scenario $[q^*, ½, q^\circ]$ where the q-vectors indicate life-time health status level streams, and the one-period certainty equivalent

q_t^+ for the treatment scenario $[q_t^*, \frac{1}{2}, q_t^0]$, are by additive independence equal. Also as has been noticed before, additive indepence implies indifference between the treatment scenarios $T_1=[(q_t^*, q_{t+1}^*) \frac{1}{2},(q_t^0,q_{t+1}^0)]$ *and* $T_2=[(q_t^*, q_{t+1}^0),\frac{1}{2},(q_t^0, q_{t+1}^*)]$.

Intuitively, one would not expect additive independence to hold. People are generally more risk averse with respect to life-time streams than with respect to one-period streams and, therefore, it can be expected that $q_t^+ < q^+$. Moreover, mixed scenarios such as T_2 will generally be preferred to extreme scenarios such as T_1.

No direct empirical evidence exists on the appropriateness of the additive independence assumption. Torrance et al. (1982, 1992) found clear evidence that the additive multi-attribute utility function was not appropriate, that is additive independence had to be rejected, in their evaluations of neonatal intensive care of very low birth weight infants and of long term sequelae of childhood cancer. Similarly, Eriksen and Keller (1993) rejected the additive specification of the multiattribute utility function for the toxicity and efficacy of drugs.

However, the multi-attribute functions assessed in these studies are one-period functions. It might be questioned though whether, given that additive independence does not hold for one-period multi-attribute utility functions, it is reasonable to assume that it will hold for multi-period multi-attribute utility functions.

If additive independence is considered to be too strong a condition, mutual utility independence might still hold. To my knowledge mutual utility independence has not been examined with respect to preferences for health. Studies that have assessed multi-attribute utility functions for health status like the ones by Torrance et al. and Eriksen and Keller have typically assumed mutual utility independence to hold, given that without this assumption the task would become very tedious. One could hypothesize that the results of Sutherland et al. generalize to the case where preferences for uncertain lifetime health scenarios rather than preferences for certain scenarios are considered. That is, if confronted with a choice between health status level q for certain and treatment with probability p of success, it can in general not be assumed that the outcome of this choice problem is independent of the health status levels in the other periods. In case the results of Sutherland et al. do indeed carry over to preferences over uncertain outcomes, mutual utility independence is most likely to be violated in the case of relatively serious and/or chronic illnesses.

Together, additive independence, equal scaling constants and stable preferences imply a constant proportional trade-off of life years for health status. Some additional empirical evidence is available about the appropriateness of this assumption under uncertainty. McNeil et al. (1981) found that individuals were only willing to trade-off life-years against improvements in health status level (in their study a change from less than perfect speech to normal speech) if the number of remaining life-years was more

than five. Even if the number of remaining life-years was larger than 5, individuals did not behave according to the constant proportional trade-off assumption, but rather according to the principle of increasing proportional trade-off. In principle both unequal scaling constants and unstable intertemporal preferences could cause this violation. More research on whether individuals place different weights on different phases in their life-cycle and on whether standard gamble valuations of health states vary with age is necessary before anything definitive can be said.

The three QALY assumptions also impose risk neutrality with respect to life years. This is easily seen by observing that equation (12) is linear with respect to life years. McNeil et al. (1978) found risk aversion with respect to life years in a group of patients with operable bronchogenic carcinoma. The same authors also found risk aversion with respect to life years in a group of subjects from the general public in their study about the trade-offs between speech and survival (McNeil et al., 1981).

As the available evidence suggests, it should not be expected that the assumptions underlying the QALY model will in general be satisfied. Even though some of the basic QALY assumptions can in principle be relaxed, QALYs and HYEs are likely to yield different results. This will not necessarily lead to reversals of preferences though. It remains to be shown how likely these reversals are to occur in practice.

Finally, as has been outlined in sections 4 and 5, both the QALY and the HYE approach assume that preferences for health can be considered separately, that is health and non-health attributes in the individual's preference structure are independent. Empirical evidence on the appropriateness of this assumption is scarce. Viscusi and Evans (1990) have studied wage-risk trade-offs in a sample of chemical workers. They found that the utility of wealth depended on the state of health. More precisely, the marginal utility of wealth decreased with decreases in health status level. This result challenges the assumption of an overall utility function in which all attributes are utility independent of their respective complements. Rejecting this assumption means rejecting the multilinear form of the utility function. Rejecting the multilinear specification of the utility function means rejecting the multiplicative and additive specifications since these are special cases of the multilinear utility function. More research is clearly necessary, but the results by Viscusi and Evans suggest that even decision making based on HYEs may give misleading results.

3.8 Possible pitfalls of HYEs

Though, as has been argued above, a HYE will not by definition correctly represent the individual's preference structure, the HYE approach is theoretically sounder than

the QALY approach. In using HYEs, one does not have to worry about the assumptions of additive independence, equal weights for equal improvements in health status in any phase of the life cycle and stable preferences over time. HYEs are more general than QALYs and therefore more likely to reflect individual preferences. Unfortunately this greater theoretical soundness is achieved at a cost: eliciting HYEs is a rather time-consuming and complicated task. In case of a health care program yielding different possible health status outcomes and a probability distribution of survival years, the HYE task becomes very cumbersome. Mehrez and Gafni (1991) propose an approximation technique to solve this problem. However, if an approximation technique is used it can no longer be claimed that "the advantage of the HYE measure .... is that it stems directly from the individual utility function and thus fully reflects the individual's preferences" (Mehrez and Gafni, 1991). The price paid for an increase in practical feasibility is a decrease in theoretical soundness.

A further problem is associated with the use of HYEs as a societal decision rule. The HYE has been developed as an individual measure of preference. However, resource allocation decisions in health care are typically societal decisions, requiring the aggregation of the preferences of individuals. The HYE in its original form cannot address this kind of question since it gives no guidance concerning the aggregation procedure. Gafni and Birch (1991) have shown how HYEs can be made consistent with several aggregation procedures (equity algorithms as they call them). However, it should be noted that each of these equity algorithms imposes a specific type of utility function on the individuals constituting society. For example, in subscribing to the equity principle that a life in full health should be given equal weight for every member of society, society implicitly imposes the following type of utility function for health on its individuals:

$$U(Q,T) = U(Q,T^*) + U(Q,T-T^*) \tag{19}$$

where T denotes the individual's total lifetime, T^* denotes the individual's remaining lifetime and $T - T^*$ denotes the time lived thus far by the individual.

It is somewhat contradictory that an approach, which claims its superiority on the basis of imposing no restrictions whatsoever on the individual utility function for health, ends up by imposing restrictions on this utility function in order to be applicable in societal decision making concerning the allocation of health care resources. However, even as a societal decision rule, the HYE approach still allows more freedom to the individual utility function than the QALY approach.

Note that within the above adjustment algorithm there is an additional problem: the selection of an appropriate value for full lifetime T. As Gafni and Birch observe, either individuals who are not newly borns have to evaluate health states for negative

periods of time or individuals have to perceive extreme lengths of life. It is a well-known result from the literature on choice under uncertainty that reference levels exert a major influence on individual risk attitudes (Schoemaker, 1982; Tversky and Kahneman, 1991; Luce et al., 1993). Setting unrealistic reference states is likely to lead to unrepresentative results. Using QALYs does not offer an easy way out of this problem. As Gafni and Birch have shown, applying QALYs as a societal decision rule requires making equity judgements as well. If one is willing to accept the equity principle that a life in full health should count equally for every individual, this requires, as in the HYE model, the specification of a reference state.

3.9 A compromise between QALYs and HYEs

In the previous sections it has been argued that both QALYs and HYEs have their pros and cons. QALYs are only under fairly restrictive assumptions equal to utilities, but are easy to measure. HYEs on the other hand more closely reflect the individual's preference structure, though without being an exact representation of it in every situation. This greater theoretical soundness is achieved at a price: a more involved measurement procedure. Moreover, applying the HYE model as a societal decision rule involves sacrificing some theoretical soundness. It would be appealing to have a measure that combines the advantages of both measures, while at the same time avoiding their disadvantages. In this section an index is proposed that to some extent attains this rather ambitious goal.

The most restrictive assumption of the QALY-procedure seems to be additive independence of the health states in different periods. If additive independence is replaced by the weaker condition of mutual utility independence, the utility function for health becomes:[10]

$$u(q) = (1/k)^* \left\{ \prod_{t=1}^{T} [1 + k k_t u_t(q_t)] \cdot 1 \right\} \tag{20}$$

where

1. $u(q)$ and $u_t(q_t)$ have been normalized
2. $k_t = u(q^*_t, q^{c\,0}_t)$ with q^c_t denoting the complement of q_t and

[10] For a proof of this result see Keeney and Raiffa (1976).

$$\sum_{t=1}^{T} k_t \neq 1 \tag{21}$$

3. k is a scaling constant that is a solution to

$$1 + k = \prod_{t=1}^{T} [1 + k k_t \, u_t(q_t)] \tag{22}$$

Equivalently, (20) can be written as

$$k u(q) + 1 = \prod_{t=1}^{T} [1 + k k_t \, u_t(q_t)] \tag{23}$$

Recall that utility functions are unique up to a positive linear transformation. Thus, if $k > 0$ then $u^+(q) = k u(q) + 1$ and $u_t^+(q_t) = 1 + k k_t u_t(q_t)$ are also utility functions and

$$u^+(q) = \prod_{t=1}^{T} u_t^+(q_t) \tag{24}$$

If $k < 0$ then $u^+(q) = \cdot [k u(q) + 1]$ and $u_t^+(q_t) = \cdot [1 + k k_t u_t(q_t)]$ are utility functions so again

$$u^+(q) = \prod_{t=1}^{T} u_t^+(q) \tag{25}$$

So even when additive independence is relaxed to mutual utility independence, the procedure to calculate the number of utilities associated with a health care program is still rather simple.

One problem remains: the assessment of the scaling constants (k and the various k_t's). This task can be greatly simplified though by imposing the (convenience) assumption of equal k_t's for every phase in the life cycle. Under this assumption, the easiest way to proceed is by determining k_1 from the following standard gamble question: determine p^* such that indifference holds between $(q_1^*, q_2^0, \ldots, q_T^0)$ and the treatment option $[(q_1^*, q_2^*, \ldots, q_T^*), \; p^*, (q_1^0, \; q_2^0, \ldots, q_T^0)]$. That is compare the certainty of one year in full health with a treatment option which offers a probability p^* of success (full health for the rest of life) and a probability $1 - p^*$ of failure (immediate death). The indifference value p^* is equal to k_1 which is equal to all other k_t's by

assumption. Once the T values of k_t are known, k can be solved from equation (22). This gives all the information necessary to calculate u^+ from (24).

The proposed aggregation procedure depends on the adopted equity principle (Gafni and Birch, 1991). For example, if it is accepted that one healthy year should count equally for each individual, the procedure is as follows. Set the difference between $u(q_t^*)$ and $u(q_t^0)$ equal to 1 for every individual in each time period and aggregate these individual values into societal values. Similarly aggregate the various individual k_t's in a set of societal k_t's. From these the societal value of k can be calculated.

In the derivation of the above index, use has been made of two assumptions: mutual utility independence and equal values of k_t.[11] It cannot be expected that these assumptions will hold in every situation. The assumption of equal values for the different k_t's can be made less restrictive by introducing discount rates. However, as has been noted above, this will not solve all problems. Alternatively, the weights associated with different phases of the life-cycle could be directly assessed by asking more questions. For obvious reasons, this requires replacing death as the worst health state. However, even then subjects may find it hard to imagine profiles of the type $(q_1^0, q_2^0, \ldots, q_{t-1}^0, q_t^*, q_{t+1}^0, \ldots, q_T^0)$. Therefore, a reformulation of the model in a disutility format is to be recommended. Such a procedure is more likely to produce reliable answers.

Even some violations of mutual utility independence are allowed. Since health status is made up of several dimensions, e.g. mobility, pain, self-care, the attributes of the lifetime utility function for health are vectors consisting of scores on these various dimensions. That is, the various component utility functions $u_t(q_t)$ are themselves multiattribute utility functions nested within a higher level multiattribute utility function. Nesting multiattribute utility functions provides additional degrees of freedom, which permit trade-offs between two attributes to depend on other attributes. This allows for some violation of the mutual utility independence assumption.

Stable intertemporal preferences can but need not necessarily be assumed. Calculating utilities for various age groups and testing whether these are significantly different seems the appropriate procedure before stability of preferences over time can be assumed.

---

[11] It is also still assumed that the axioms of expected utility theory hold

3.10 Concluding remarks

It is not claimed that the above index is the ideal one. Still some restrictive assumptions have been made even though the implications of these assumptions can be relaxed by discounting, reformulation of the model in a disutility format and by nesting the component one-attribute utility functions. The above index attempts to combine the advantages of using QALYs (easy to calculate) with those of HYEs (theoretically sound at the individual level). The above index is easier to calculate than HYEs though not as easy as QALYs given that extra questions have to be asked to determine the k_i's. The index more closely approaches the individual utility function than the QALY model, given that less restrictive asumptions have been imposed on the individual's preference structure.

Problems arise when health status levels worse than death have to be evaluated. In such cases mutual utility independence is violated. Though nesting may solve some problems, extreme care should be taken in applying the index to evaluate health status levels worse than death. On the other hand, in such cases the QALY procedure cannot be applied either and the existence of HYEs is not guaranteed as Mehrez and Gafni (1991) show. How best to handle health outcomes worse than death remains an important issue on the research agenda for outcome measurement in health. At the moment, the recommended procedure is to use several measures and to test extensively for the sensitivity of the results obtained.

The step from QALYs to HYEs implied relaxing several restrictions that had been imposed on the individual's preference structure. The above index is in fact a step backwards on the road between QALYs and HYEs, since a restriction (mutual utility independence) has been re-introduced. An alternative course would be to take the opposite direction and investigate the implications of relaxing even more assumptions. This would lead to a sort of Grossman formulation in which the individual's overall utility function can no longer be assumed to be separable. Exploration of this road is an interesting topic for future research.

Appendix

Theorem 3.2 can be proved by combining theorems 5.1 and 6.4 in Keeney and Raiffa (1976), which are based on results derived by Fishburn (1965, 1970).

Consider first an individual who only lives two periods. Given additive independence this individual will be indifferent between the treatment scenarios $[(q_1, q_2), ½, (q_1^0, q_2^0)]$ and $[(q_1, q_2^0, ½, (q_1^0, q_2)]$ since they have the same marginal probability distribution on quality of life levels in the two periods.

Equating expected utilities and setting $u(q_1^0, q_2^0) = 0$, which is allowed given freedom to scale the utility function gives $½u(q_1, q_2) = ½\ u(q_1, q_2^0) + ½\ u(q_1^0, q_2)$.

Defining $u(q_1, q_2^0) = k_1\ u_1(q_1)$ and $u(q_1^0, q_2) = k_2\ u_2(q_2)$ to allow for free scaling of the single period utility functions gives the additive form of the multi-attribute utility function.

The generalization to the T period case is straightforward. Define Y as $\{q_2, \ldots, q_T\}$. Then from the above $u(Q_T) = k_1 u_1(q_1) + k_y u_y(q_2, \ldots, q_T)$.

Break down u_Y by defining Z as $\{q_3, \ldots, q_T\}$. Apply the above again to yield:

$u_y(q_2, \ldots, q_T) = k_2 u_2(q_2) + k_z u_z(q_3, \ldots, q_T)$.

Proceeding this way and substituting the obtained expressions in each other gives the additive multi-attribute utility function.

That the additive multi-attribute utility function implies additive independence can immediately be derived by calculating the expected utilities of the treatment scenarios.

4

An experimental test of constant proportional trade-off and utility independence[1]

Summary

Pliskin, Shepard and Weinstein (1980) have identified three preference conditions that ensure that quality-adjusted life years (QALYs) represent preferences over chronic health profiles. This chapter presents an experimental test of the descriptive validity of two of these preference assumptions: utility independence (UI) and constant proportional trade-off (CPT). The results of our experiment provide support for CPT: both within subjects and between subjects we could not reject CPT. The results are less supportive for UI. Within subjects UI was rejected. Between subjects we could not reject UI, but this may be due to a lack of statistical power. Analysis of the individual responses reveals that without adjustment for imprecision error 39 respondents (22.8%) satisfy CPT. Twenty-three respondents (13.4%) satisfy UI. Adjusted for imprecision error, 155 respondents (90.1%) satisfy CPT and 130 respondents (75.6%) satisfy UI. Pliskin, Shepard and Weinstein have further derived that if an individual's preferences satisfy both CPT and UI, then these preferences can be represented by a more general, risk-adjusted QALY model. Without adjustment for imprecision error 10 respondents (5.8%) satisfy both CPT and UI. Adjusted for imprecision error 118 respondents (68.6%) satisfy both CPT and UI. The results of this chapter reveal a strong impact of adjustment error. This strong impact suggests that CPT and, to a lesser extent, UI hold approximately. Moreover, adjusted for imprecision error, a majority of respondents satisfy both CPT and UI. This implies that the use of a risk-adjusted QALY model provides a reasonable description of individual preferences in medical decision contexts.

[1] Based on Bleichrodt, H. and M. Johannesson, "An experimental test of constant proportional trade-off and utility independence," Medical Decision Making (accepted for publication).

4.1 Introduction

In health care, as in other areas of social policy, decisions have to be made concerning the allocation of scarce resources. Cost utility analysis in which quality-adjusted life years (QALYs) are used as outcome measure is intended to guide health care policy making. Over the last decade, QALY based decision making has become increasingly popular. QALYs provide a straightforward way to combine the two main outcomes of health care programs, quantity of life and quality of life, into one single index measure. A further advantage of using QALYs is that they have intuitive appeal. However, ever since the introduction of QALYs, their theoretical properties have been a matter of concern. Pliskin, Shepard and Weinstein (1980) were the first to provide an axiomatic analysis of QALYs. These authors show that, given an individual preference relation over gambles involving quantity of life and constant quality of life that satisfies the axioms of expected utility theory, three conditions have to be imposed on this preference relation to ensure that it can be represented by the QALY model. These conditions are referred to as "(mutual) utility independence," "constant proportional trade-off," and "risk neutrality on life years." Pliskin et al. have further derived that imposing utility independence and constant proportional trade-off, but not risk neutrality on life years, ensures that the individual preference relation can be represented by a general QALY model in which life years do not enter linearly, but are adjusted for risk attitude.

Identifying the preference conditions on which the QALY model depends, allows an assessment of both the extent to which it is rational for an individual to behave according to the model (i.e. the normative validity of the model) and the extent to which the model actually describes individual preferences (i.e. the descriptive validity of the model). Most of the available empirical evidence on the descriptive validity of the QALY model is about risk neutrality on life years. The majority of these studies rejects risk neutrality on life years. An exception is a study by Miyamoto and Eraker (1985) in which risk neutrality is found to hold for the "average respondent." The little empirical evidence that is available to date on constant proportional trade-off and utility independence is inconclusive. Loomes and McKenzie (1989) argue that these conditions cannot be expected to hold in every decision situation and are not supported by empirical evidence. On the other hand, Miyamoto and Eraker (1988) provide empirical evidence that supports utility independence of quantity of life from quality of life.

A disadvantage of the characterization of the QALY model by Pliskin et al. is that it only applies to health profiles of a constant quality. More general characterizations of the QALY model exist that allow quality of life to vary over

time [Broome, 1993; Bleichrodt, 1995]. However, a major advantage of the characterization by Pliskin et al. is that two of their preference conditions are directly related to two commonly used methods of estimating quality weights for health states: the standard gamble and the time trade-off. We explain in the next section that the condition of utility independence (UI) is related to the standard gamble and that the condition of constant proportional trade-off (CPT) is related to the time trade-off. These relationships allow straightforward tests of utility independence and constant proportional trade-off in an experimental design. The present study reports evidence from an experiment aimed at testing the descriptive validity of constant proportional trade-off and utility independence of quality of life from quantity of life. We both provide evidence on the extent to which the respondents in our experiment satisfy each of these conditions separately and on the extent to which they satisfy both these conditions simultaneously. Given that previous studies have in general rejected risk neutrality on life years and thereby have rejected the descriptive validity of the QALY model, it is interesting to examine whether the risk adjusted QALY model proposed by Pliskin et al. provides a better description of individual preferences. As observed above, this model depends on utility independence and constant proportional trade-off. Therefore the results of our study allow to draw inferences with respect to the descriptive validity of the risk adjusted QALY model.

The structure of the chapter is as follows. In the next section we explain in more detail the theory of QALYs that we briefly touched upon in this introduction. Then experimental methods and results are discussed. The chapter finishes with a discussion of the results and of the implications of this chapter for the use of QALYs in medical decision making.

4.2 Theoretical analysis of QALYs

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

1944]. Then a real-valued, cardinal, utility function $U(Q,T)$ exists, the expected value of which represents individual preferences over gambles involving quality of life and quantity of life.

Pliskin et al. have derived that QALYs are a valid von Neumann Morgenstern utility function if in addition to the von Neumann and Morgenstern axioms three other conditions are imposed on the individual preference relation: (mutual) utility independence, constant proportional trade-off and risk neutrality on life years. We consider each condition in turn.

4.2.1 Utility independence

When we fix one of the two attributes in the utility function over health profiles at a particular value, utility independence imposes that preferences with respect to gambles over the other attribute do not depend on the particular value chosen. Formally, utility independence implies that $[p_1,(Q_1,T_1);\ldots\ldots;p_n,(Q_n,T_1)]$ is preferred to $[r_1,(Q_1,T_1);\ldots;r_n,(Q_n,T_1)]$ if and only if $[p_1,(Q_1,T_2);\ldots\ldots;p_n,(Q_n,T_2)]$ is preferred to $[r_1,(Q_1,T_2);\ldots;r_n,(Q_n,T_2)]$ for all T_1, T_2. A similar expression holds when Q is held fixed and T varies. Denote by $W(Q)$ a utility function over quality of life and by $V(T)$ a utility function over quantity of life. Keeney and Raiffa (1976) have shown that utility independence implies that $U(Q,T)$ is either multiplicative, i.e. $W(Q)*V(T)$, or additive, i.e. $W(Q)+V(T)$.[2] The additive model depends on a condition Pliskin et al. refer to as "marginality." Pliskin et al. and Miyamoto and Eraker (1985,1988) provide arguments why the additive model is not realistic in the medical context. The additive model can be excluded by adding an entirely plausible condition to the model: that for a time duration of zero life years the individual is indifferent between all quality of life levels.[3] If the additive model is discarded, utility independence can be shown to imply: $U(Q_1,T_1)/U(Q_2,T_1)=U(Q_1,T_2)/U(Q_2,T_2)$.[4]
If we plot $U(Q,T)$ against T, holding quality of life fixed, then utility independence guarantees that the shape of $U(Q,T)$ is the same regardless of the level at which

---

[2] As one of the referees reminded us, this only holds when V and W are rescaled in line with U. For more details see Keeney and Raiffa [p.289-291].

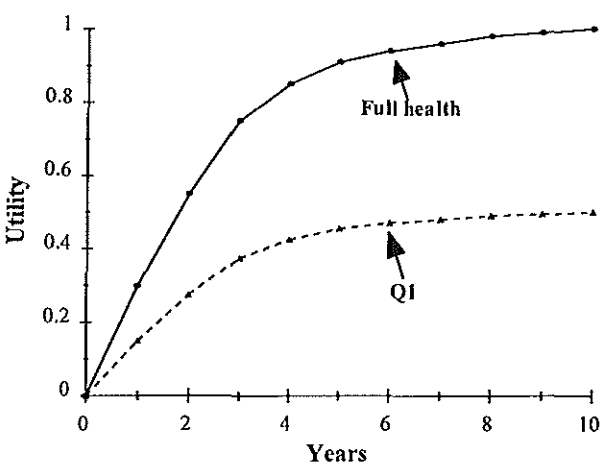
[3] A proof of this result has been provided in chapter 2.

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

quality of life is held fixed. This is illustrated in figure 4.1 for life durations up to 10 years, where, for convenience, full health is selected as quality of life level Q_2 . If utility independence holds, then for all health states the fraction of the utility of full health is independent of the time horizon. In figure 4.1 for instance the utility of health state Q_1 is 0.5 of the utility of full health for all time horizons.

Utility independence facilitates the determination of standard gamble quality weights. The standard gamble determines the quality weight of a health state by comparing a specific number of years in this health state to a gamble with a probability *(p)* of the same number of years in full health and a complementary probability *(1-p)* of immediate death. The probability of full health *(p)* is varied until the individual is indifferent between the alternatives. Suppose *W(Q)* is scaled such that *W(full health) = 1* and *W(death) = 0.* The quality weight of the health state is then set equal to p^*, where p^* is the probability for which the individual is indifferent. Thus, the standard gamble measures the utility of a health state as the fraction of the utility of full health. By utility independence, this fraction does not depend on the number of years the measurement is carried out for. The only restriction is that the number of life years be equal for the certain health state and for full health. In figure 4.1 the standard gamble quality weight for Q_1 is equal to 0.5 regardless of the time horizon the assessment is carried out for.

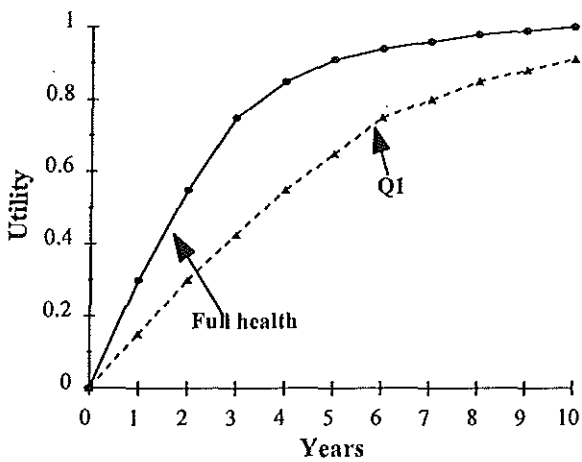
Figure 4.1: The utility function for life-years under utility independence



4.2.2 Constant proportional trade-off

Constant proportional trade-off imposes that if an individual is indifferent in a choice between T years in health state Q_1 and αT years $(0 \leq \alpha \leq 1)$ in a more attractive health state Q_2, then this individual should also be indifferent between βT years $(\beta \geq 0)$ in Q_1 and $\beta\alpha T$ years in Q_2. The proportion "years in Q_1 divided by years in Q_2" is constant by the condition of constant proportional trade-off (in the above choice situation this proportion is equal to $1/\alpha$). Constant proportional trade-off is illustrated in figure 4.2. Figure 4.2 displays the situation where the individual is willing to sacrifice 50% of his remaining life span in Q_1 to improve his health to Q_2, which is set equal to full health for convenience. By constant proportional trade-off this proportion holds for any time horizon. Thus, as can be seen from figure 4.2, the individual is indifferent between 6 years in Q_1 and 3 years in full health, but also between 4 years in Q_1 and 2 years in full health.

Figure 4.2: The utility function for life years under constant proportional trade-off



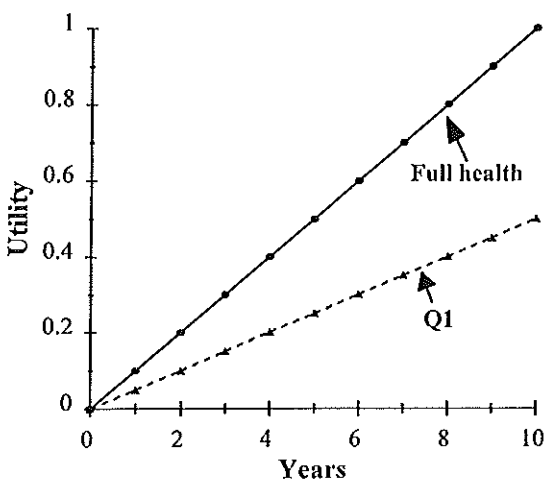
Constant proportional trade-off facilitates the assessment of time trade-off quality weights. The time trade-off determines the quality weight of a health state by comparing T years in the health state to X years in full health. The number of years in full health (X) is varied until the individual is indifferent between the

alternatives. The quality weight of the health state is then set equal to X/T. The time trade-off thus measures quality weights as the equivalent fraction of healthy years. By the assumption of constant proportional trade-off this fraction will be independent of the time horizon the assessment is carried out for.

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

4.2.3 Risk neutrality on life years

Figure 4.3: The utility function for life years under risk neutrality



The only situation in which standard gamble and time trade-off will, at least in theory, elicit identical quality weights is the situation in which $r=1$, i.e. the situation in which the individual is risk neutral with respect to life years. This situation characterizes the QALY measure most frequently used in cost utility analysis. Risk neutrality with respect to life years implies a utility function for life years that is linear in life years. Figure 4.3 illustrates a risk neutral utility function both for full health and for a health state Q_1. In figure 4.3 both the time trade-off weight and the standard gamble weight are equal to 0.5.

4.2.4 Empirical evidence

Empirical evidence on the QALY assumptions is fairly scarce. In a review of the literature available to that date, Loomes and McKenzie (1989) drew rather negative conclusions. However, Loomes and McKenzie only examined whether the conditions hold exactly. They did not allow for a certain imprecision in respondents' preferences. Miyamoto and Eraker (1985) tested utility independence of quantity of life from quality of life. Their results support this condition. With respect to the trade-off between quality of life and quantity of life, several authors have observed that increasing proportional trade-off (i.e. an individual is willing to sacrifice relatively more remaining life-years when the number of remaining life years increases) is more consistent with individual preferences than constant proportional trade-off [Pliskin et al., 1980; McNeil et al., 1981]. Moreover, in some studies it was observed that for small time durations individuals are not willing to trade off any life years for an improvement in quality of life [McNeil et al., 1981; Miyamoto and Eraker, 1988]. Apparently individual preferences are lexicographic for small time durations: individual choices are fully determined by the number of life years. Stiggelbout et al. (1994) tested whether under risk adjustment time trade-off weights are equal to standard gamble weights, as predicted by the risk adjusted QALY model. Their general finding is that risk adjustment does indeed exert a converging influence. Some more evidence is available on risk neutrality. First, several studies have shown that time trade-off and standard gamble do elicit different quality weights where the risk neutral QALY model predicts equality. However, one should be careful in drawing inferences for risk neutrality from these findings. It is well known that the way the standard gamble question is asked in health utility assessment, i.e. by probability equivalence, exerts an upward bias on the elicited utilities [cf. Hershey and Schoemaker, 1985]. Second, studies that directly tested risk neutrality on life years by assessing utility functions over life

years, typically reject risk neutrality [Stiggelbout et al., 1994; McNeil et al., 1978; Verhoef et al., 1994; Maas and Wakker, 1994]. The one exception is the study by Miyamoto and Eraker (1985) mentioned in the introduction.

Below we report the results of an experiment aimed at testing constant proportional trade-off and utility independence of quality of life from quantity of life. To the best of our knowledge the latter condition has not been tested before. Miyamoto and Eraker (1988) tested the converse: whether quantity of life is utility independent from quality of life. The importance of testing whether constant proportional trade-off and utility independence hold simultaneously follows from the rejection of risk neutrality on life years in various studies. The results of these studies challenge the descriptive validity of the risk neutral QALY model. The question then emerges whether the more general risk-adjusted QALY model performs better in describing individual preferences. As explained above, the risk-adjusted QALY model is characterized by constant proportional trade-off and utility independence. Testing these two conditions provides insight in the descriptive validity of the risk adjusted QALY model.

Testing constant proportional trade-off and utility independence separately is also interesting in its own right. Constant proportional trade-off ensures that time trade-off weights are independent of the time horizon the assessment is carried out for. Utility independence ensures that standard gamble weigths are independent of the time horizon the assessment is carried out for. Utility independence also allows the estimation of another generalization of the risk neutral QALY model, in which risk neutrality and constant proportional trade-off are relaxed [cf. chapter 3].

4.3 Methods

4.3.1 Respondents

Eighty students at the Stockholm School of Economics and 92 students at Erasmus University Rotterdam took part in the experiment. The students were all undergraduates and were recruited from courses in economics, statistics and health policy. The students were paid approximately $15 for their participation. The experiment was carried out in 17 sessions lasting approximately one hour with on average ten respondents per session. The procedure followed in each session was to explain first a specific task to respondents, then to ask respondents to perform the specific task and then to explain the next task. A "master" version of the experiment

was designed in English. This "master" version was subsequently translated into Swedish and Dutch. Before drafting the final version, we tested the questionnaire extensively both in Stockholm and in Rotterdam using faculty staff members as respondents.

4.3.2 Health states

We selected eight health states to be included in the questionnaire. The health states were taken from the Maastricht Utility Measurement Questionnaire, a slightly adapted version of the McMaster Health Utility Index [Bakker et al., 1994; Rutten-van Mölken et al., 1995]. The selected health states correspond to commonly occurring types of back pain and rheumatism. The health states consist of four dimensions: general daily activities, self care, leisure activities and pain. The health states were indicated by capital letters and were described on a set of cards, which were handed out to respondents at the beginning of each session. Health states B and D are relevant for the analysis of this chapter. They are described in table 4.1. Health state B is clearly more attractive than health state D.

Table 4.1: Health states B and D

B

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

D

- Unable to perform some tasks at home and/or at work.
- Able to perform all self care activities (eating, washing, dressing) albeit with some difficulties.
- Unable to participate in many types of leisure activities.
- Often moderate to severe pain and/or other complaints.

4.3.3 Questionnaire

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

In the third section constant proportional trade-off was tested. Time trade-off quality weights for health states *B* and *D* were determined. Respondents were encouraged to indicate first the values of *X*, the number of healthy years, for which they definitely preferred to be in health state *B* and *D* respectively, then the values of *X* for which they definitely preferred to be in full health and finally those values of *X* for which they found it hard to choose between the alternatives. Respondents were explicitly told that all profiles would result in death after the indicated number of years. The general introduction to the time trade-off questions can be found in appendix 1. It was pointed out to respondents both in the text and in the oral explanantion of the task that they could indicate a range of values for *X* for which they found it hard to choose between the alternatives. The response strategy we suggested to respondents is likely to lead to an interval of indifference values. When respondents first mark all the values for which they have clear preference for one of the alternatives, they end up with a range of values for which they are not certain which alternative to prefer. We told respondents to mark these values with the symbol for indifference. This format was adopted to allow for the fact that respondents are likely to have imprecise preferences [Dubourg et al., 1994]. Respondents are unfamiliar both with the health states to be assessed and with the idea of trading off life-years. As a result their preferences may be imprecise. In our format we attempted to take this imprecision into account. If respondents indicated a range of values for which they could not choose between the options, we interpreted this range of values as their personal confidence interval (*PCI*).

Personal confidence intervals should not be confused with statistical confidence intervals. However, the idea behind them is somewhat similar. Personal confidence intervals indicate a range of values for which an individual expresses indifference. That is, all these values cannot be distinguished from the "true"

indifference value. The interpretation of personal confidence intervals we used, is similar to the interpretation of statistical confidence intervals: if two values differed, but the personal confidence intervals in which these values were contained, overlapped, we interpreted the difference between the values as not being significant in terms of the individual's personal confidence interval. If the personal confidence intervals did not overlap, the difference was interpreted to be significant.

In the results section, we present the responses both with and without adjustment for imprecision of preferences. For those respondents who had indicated a personal confidence interval we used this interval to adjust for imprecision error. For those respondents who had not indicated a personal confidence interval, we constructed an artificial personal confidence interval by adjusting for the median imprecision error. The median imprecision error was computed from the responses of those respondents who had indicated a personla confidence interval. For comparison we present in appendix 3 the results when no artificial personal confidence intervals are constructed. These results interpret the responses of those respondents who did not indicate a personal confidence interval as being precise. It should be emphasized here that we find it hard to believe that these respondents were indeed certain of their responses, given the unfamiliarity of the health states and of the tasks they were requested to perform,.

Both versions of the questionnaire consisted of three time-trade-off questions. Version 1 started with a time trade-off question for 10 years in health state *D* (*D10*), followed by a time trade-off question for 30 years in health state *D (D30)* and a time trade-off question for 30 years in health state *B (B30).* Version 2 started with a time trade-off question for 10 years in health state B *(B10),* followed by a time trade-off question for 30 years in health state B and a time trade-off question for 30 years in health state D. This setup allowed us to test whether constant proportional trade-off holds at the individual level. For the respondents answering a version 1 questionnaire we compared the answers to questions *D10* and *D30.* For the respondents answering a version 2 questionnaire we compared the answers to questions *B10* and *B30.*

Two types of biases may have affected the responses that we obtained. A first bias may occur as a result of asking the time trade-off questions for 10 and 30 years immediately after one another. This setup may cause responses to be anchored. For example, respondents may adopt a proportional heuristic [Stalmeier et al., 1995] in answering the questionnaire. That is, they simply state a fixed percentage of the remaining lifetime, even though their preference relation does not actually satisfy constant proportional trade-off. To examine the possibility of an anchoring strategy we included a between subjects test of constant proportional trade-off, which is not

affected by anchoring: the mean time trade-off weight for *B30* in version 1 was compared with the mean time trade-off weight for *B10* in version 2. Similarly, the mean time trade-off weight for *D30* in version 2 was compared with the mean time trade-off weight for *D10* in version 1. Apart from this test, some inferences with respect to the effect of anchoring can be drawn from a comparison between the two versions of the answers to *B30* and *D30*. For example, under an increasing proportional trade-off one would expect *D30* to be higher for version 1, in which *D10* was also included. If the individual preference relation satisfies increasing proportional trade-off *D10* will be higher than *D30*. However, by anchoring we would expect *D10* and *D30* to be equal. Therefore if *D10* is asked first, given increasing proportional trade-off, anchoring will induce an upward bias on *D30*.[5] Similarly, under increasing proportional trade-off *B30* can be expected to be higher in version 2. This test assumes a random distribution of preferences in the two samples. We have no reason to believe that this assumption does not hold. Respondents were allocated randomly to versions and we have no indication that significant bias was introduced by the allocation process.

The second bias may have been introduced by the fact that the versions differed in the six health states that were evaluated. As explained above, version 1 contained more severe health states than version 2. This may have made health states *B* and *D* appear more attractive, and may thus have resulted in higher weights for health states *B* and *D* in version 1. Two points are worth making with respect to this possible bias. First, it will only affect our between subjects tests. Within subjects obviously the same version was used and we can still compare the answers. Second, if this bias would indeed affect our results, we would expect it to be stronger for health state *B* than for health state *D*. The reason for this is that health state *D* was in both versions still the worst health state. This was reflected by the ranking exercise: all but two (version 2) respondents ranked health state *D* as the worst health state. Analysis of the rating scale valuations confirmed this expectation. The rating scale valuation for health state B differed significantly across versions, but the difference between the rating scale valuations was not significant for health state *D*. Therefore if the two versions differed significantly in the time trade-off weight for health state *B*, but to a smaller extent in the time trade-off weight for health state *D*, we interpret this response pattern as an indication that the inclusion of different health states has produced a bias in responses.

---

[5] Obviously, under decreasing proportional trade-off and anchoring the opposite pattern holds: D30 will be lower in version 1.

In section 4 utility independence of quality of life from quantity of life was tested.[6] Standard gamble quality weights for health states *B* and *D* were determined. Again respondents were explicitly informed that all health profiles would be followed by death. Probability elicitation was by means of a line of values for the probability of successful treatment (full health). Next to this line a line was drawn with the complementary probability of failure of treatment (immediate death). This display was chosen in an attempt to control for a potential framing bias: only displaying the probability of successful treatment might induce an individual to focus on successful treatment, not sufficiently taking into account the probability of failure of treatment. Psychological evidence on the influence of reference effects on choice is abundant [e.g. Kahneman and Tversky, 1979]. Similar to the time trade-off question, an attempt was made to take imprecision of preferences into account. First, respondents were asked to indicate those values of *p*, the probability of successful treatment, for which they definitely preferred the certain option, then those values of *p* for which they definitely preferred the treatment option (gamble), and finally those values of *p* for which they found it hard to choose between the options. The general explanation of the standard gamble questions can be found in appendix 2. Again it was pointed out to respondents both in the description of the task and in the oral explantion, that they were allowed to indicate a range of values of *p* for which they found it hard to choose between the options. This range of values was then interpreted as the personal confidence interval for *p*.

Like section 3, section 4 consisted of two versions.. In version 1 the order in which the questions were asked was *D10, D30, B30*. In version 2 the order in which the questions were asked was *B10, B30, D30*. For version 1 respondents utility independence was tested by comparing their answers to the two standard gamble questions that involved health state *D*. For version 2 respondents utility independence was tested by comparing their answers to the two standard gamble questions that involved health state *B*. To allow for possible anchoring effects we compared across versions the answers to *D10* and *D30* and to *B30* and *B10*. Finally we compared *B30* and *D30* across versions to get an impression whether the results may have been affected by either anchoring bias or a bias introduced by the difference in included health states between the versions.

Sections 5 and 6 consisted of two tasks that are not relevant for the present analysis, but will be reported elsewhere.

---

[6] In the sequel of the chapter we will for convenience speak about utility independence when we mean utility independence of quality of life from quantity of life.

4.4 Statistical analysis

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

Given the size of our sample, at a significance level of 5%, the paired t-test is able to detect a difference of 0.25 times the standard deviation with a power of over 90%. Given that standard deviations for time trade-off and standard gambles quality weights reported in the literature rarely exceed 0.2,[7] the probability of detecting a true difference of 0.05 by the paired t-test is higher than 90%. The power of the independent samples t-test to detect a difference of 0.25 times the standard deviation at a significance level of 5% is 45%. The power to detect a difference of 0.5 times the standard deviation is higher than 90%. In comparison, in the study by Miyamoto and Eraker (1988) the median probability to detect a difference of 0.05, given a significance level of 5%, was estimated to be 28%. The probability to detect a true difference of 0.10 was estimated to be just over 70% in their study.

Hypotheses with respect to proportions were tested by calculating χ^2 values from the resulting 2x2 contingency tables, which were compared with 1 degree of freedom. Continuity corrections were made both in the case where proportions come from the same population and in the case where proportions come from different populations. The method used to test hypotheses with respect to proportions uses the continuous Normal distribution as an approximation to the discrete Binomial distribution. The Normal distribution corresponds better to the Binomial distribution when a correction is made to the observed frequency to allow for the fact that variables can only take integer values [Altman, 1991].

To examine whether a systematic relationship exists between satisfying contant proportional trade-off and satisfying utility independence, we used a binary choice model. We took the 0-1 variable "satisfying constant proportional trade-off yes/no" as the variable to be explained. This variable takes the value 0 if a respondent does not satisfy constant proportional trade-off and 1 if a respondent

[7] Most standard deviations in our study were also lower than 0.2

does satisfy constant proportional trade-off. The 0-1 variable "satisfying utility independence yes/no" was taken as the explanatory variable. This variable takes the value 0 if a respondent does not satisfy utility independence and 1 if a respondent does satisfy utility independence. For this analysis we could choose between a probit model, in which the error terms are distributed according to the standard normal distribution, and a logit model, in which the error terms are distributed according to the logistic distribution. Because we estimated a univariate dichotomous model, it is hard to distinguish between the two methods [Amemiya, 1994]. However, the logistic distribution has slightly heavier tails and because we could not exclude the possibility that responses to the constant proportional trade-off and utility independence questions would be concentrated in the tails we decided to use the logit model. Model performance was assessed by the Likelihood Ratio test.

4.5 Results

4.5.1 Analysis of individual responses

Constant proportional trade-off

Table 4.2 shows the results of the test of 'constant proportional trade-off on the basis of the individual data. Individuals whose choices exactly satisfy constant proportional trade-off are in category *C*. The proportions of respondents in category *C* are 18.4% and 27.1% in version 1 and version 2 respectively. The difference between these proportions is not significant. The proportion of respondents in category *C* is slightly distorted for version 2, because three version 2 respondents were not willing to trade-off any life years at all for an improvement in health. Excluding these respondents leaves 24.4% of the version 2 respondents in category *C*. Contrary to previous studies [McNeil et al., 1981], table 4.2 shows no indication that increasing proportional trade-off is a more common response pattern than decreasing proportional trade-off.

Table 4.2: Number of time trade-off responses per category

Sample	Version	n	A	B	C	D	E
Total	1	87	5	31	16	30	5
Total	2	85	5	27	23	28	2
Swedish	1	40	2	19	3	13	3
Swedish	2	40	1	13	10	15	1
Dutch	1	47	3	12	13	17	2
Dutch	2	45	4	14	13	13	1

Note: A = weight 10 years > weight 30 years and no overlap PCIs.
B = weight 10 years > weight 30 years, but overlap PCIs
C = weight 10 years = weight 30 years
D = weight 10 years < weight 30 years, but overlap PCIs
E = weight 10 years < weight 30 years and no overlap PCIs

The results presented above do not take into account that respondents' preferences are likely to be somewhat imprecise. As explained in the methods section, for those respondents who indicated a personal confidence interval we used this interval to examine overlap. However, a confidence interval was given for only 92 responses.[8] The artificial confidence interval we estimated on the basis of the median imprecision error[9] resulted in a personal confidence interval of $[TTO - 0.075; TTO + 0.075]$.

Respondents who are in categories *B* and *D* have overlapping personal confidence intervals. For these respondents the difference between the TTO valuations is interpreted as not being significant. Therefore these respondents are counted as satisfying constant proportional trade-off. Respondents who are in categories *A* and *E* have non-overlapping personal confidence intervals. For these respondents the difference between the TTO valuations is counted as being significant and their choices violate constant proportional trade-off even after adjustment for imprecision error.

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

[9] Adjusting for the mean imprecision error resulted in slightly larger personal confidence intervals: $[TTO-0.09; TTO+0.09]$. However, using the mean imprecision error to construct artificial personal confidence intervals hardly affected the results: both in version 1 and in version 2 one individual no longer violated constant proportional trade-off with imprecision adjustment.

Table 4.2 shows that the overwhelming majority of time trade-off responses satisfied constant proportional trade-off with imprecision adjustment: 88.6% in version 1 and 91.7% in version 2. The difference between proportions in the two versions is not significant.The differences between the proportions satisfying increasing proportional trade-off and decreasing proportional trade-off are not significant in both versions. For comparison, table A2 (appendix) shows that if responses are only adjusted partially for imprecision error (i.e. no artificial personal confidence intervals are constructed) 37.5% of the respondents in version 1 and 36.5% of the respondents in version 2 satisfy constant proportional trade-off.

Table 4.2 also displays that the results for the Swedish and Dutch samples are not significantly different. The only exception is the proportion of version 1 respondents who exactly satisfy constant proportional trade-off. This proportion is significantly higher in the Dutch sample ($\chi^2(1) = 6.34$; $p = 0.024$).

Utility independence

Table 4.3: Number of standard gamble responses per category

Sample	Version	n	A	B	C	D	E
Total	1	87	16	36	8	17	10
Total	2	85	12	44	15	10	4
Swedish	1	40	9	18	4	6	3
Swedish	2	40	7	20	7	5	1
Dutch	1	47	7	18	4	11	7
Dutch	2	45	5	24	8	5	3

Note: A = weight 10 years > weight 30 years and no overlap PCI's.
B = weight 10 years > weight 30 years, but overlap PCI's
C = weight 10 years = weight 30 years
D = weight 10 years < weight 30 years, but overlap PCI's
E = weight 10 years < weight 30 years and no overlap PCI's

Table 4.3 shows that overall the proportion of respondents who exactly satisfy utility independence, the respondents in category C, is equal to 13.4%. This proportion is lower than the proportion of respondents who exactly satisfy constant proportional trade-off. However, the difference is not statistically significant. The proportion of respondents in category C is higher in version 2,

17.6% versus 9.2 % in version 1. Again the difference is not statistically significant. The proportion of respondents in categories *A* and *B* is significantly higher than the proportion of respondents in categories *D* and *E* ($\chi^2(1) > 10.8$; $p = 0.000$ in both versions). This indicates that the utility of health states *B* and *D* as a fraction of the utility of full health decreases with the time horizon the assessment is carried out for. Utility independence predicts that these fractions should be constant as we have explained in the section on the theoretical properties of QALYs.

The artificial personal confidence interval constructed for those respondents who did not indicate a personal confidence interval is equal to: $[SG-0.05;SG+0.05]$.[10] Table 4.3 shows that after adjustment for imprecision error a majority of the respondents satisfies utility independence. However, the proportion of respondents who satisfy utility independence (75.6%) is still lower than the proportion of respondents who satisfy constant proportional trade-off. The difference between the proportions is highly significant ($p = 0.006$) for version 1. For version 2 this difference is only significant at the 10% level ($p = 0.06$).

It is interesting to observe from table A3 (appendix) that if only a partial adjustment is made for imprecision error, the proportion of respondents in categories *B,C,* and *D* only slightly increases to 18.0%.

No significant differences were observed between the Dutch and the Swedish sample. The pattern for the two samples separately is similar to the general pattern. The difference between the proportion of respondents who exactly satisfy utility independence and the proportion of respondents who exactly satisfy constant proportional trade-off is only significant for version 1 respondents in the Dutch sample ($p = 0.016$), the latter being higher. Adjusted for imprecision error, the proportion of respondents satisfying constant proportional trade-off is significantly different at the 10% level from the proportion of respondents satisfying utility independence for both versions in the Swedish sample ($p = 0.088$ and $p = 0.076$) and for version 1 in the Dutch sample ($p = 0.052$).

Constant proportional trade-off and utility independence

Pliskin et al. (1980) have derived that constant proportional trade-off and utility independence guarantee that individual preferences over lotteries on chronic health profiles can be represented by the risk adjusted QALY model. Table 4.4 shows to

[10] The mean imprecision error was approximately the same: 0.052. Using personal confidence intervals estimated on the basis of the mean imprecision error did not affect the results

what extent respondents satisfy both these conditions simultaneously. In table 4.4 the column *CPT+UI* shows the number of respondents who satisfy both constant proportional trade-off and utility independence when no adjustment is made for imprecision error. The column CPT_{adj} + UI_{adj} shows the number of respondents who satisfy both constant proportional trade-off and utility independence when responses are adjusted for imprecision error. Table 4.4 shows that unadjusted for imprecision error only a small proportion of respondents satisfy the two conditions simultaneously: 3.4% and 8.2% for versions 1 and 2 respectively. The difference is not statistically significant. However, adjustment for imprecision error once again has a major impact.[11] After adjustment for imprecision error a majority of respondents satisfies the risk adjusted QALY model: 62.1% and 75.3% for versions 1 and 2 respectively. The difference between the two versions is significant at the 10% level ($\chi^2(1) = 3.52$; $p = 0.063$).

Table 4.4: number of respondents satisfying constant proportional trade-off and utility independence simultaneously. Both without and with error adjustment

Sample	Version	n	CPT + MUI	CPT_{adj}+MUI_{adj}
Total	1	87	3	54
Total	2	85	7	64
Swedish	1	40	0	25
Swedish	2	40	2	31
Dutch	1	47	3	29
Dutch	2	45	5	33

Table A4 (appendix) shows that partial adjustment for imprecision error only marginally increases the proportion of respondents who satisfy both constant proportional trade-off and utility independence: the proportion rises from 5.8% to 10.5%.

Differences between the Swedish and the Dutch samples are not significant. The proportion of respondents satisfying constant proportional trade-off and utility

[11] For respondents who did not indicate a personal confidence interval we used the same artificial personal confidence intervals as before to adjust for imprecision error: $[TTO-0.075; TTO+0.075]$ and $[SG-0.05; SG+0.05]$.

independence is always higher among version 2 respondents. However, in both samples the difference is not significant.

Table 4.4 does not show whether a systematic relationship exists between constant proportional trade-off and utility independence. That is, no information is provided whether a respondent who satisfies utility independence is also more likely to satisfy constant proportional trade-off. We estimated logistic regressions to examine whether such a systematic relationship exists. Denote the probability that a respondent has a value i on the constant proportional trade-off variable given that this respondent has a value j on the utility independence variable by $P(CPT = i \mid UI = j)$. For example, the probability that a respondent satisfies constant proportional trade-off given that he satisfies utility independnece is denoted by $P(CPT = 1 \mid UI = 1)$.

Table 4.5: Results of the logistic regression estimation (no adjustment for imprecision error)

Sample	Version	P(CPT=1| UI=0)	P(CPT=1| UI=1)	Model χ^2
Total	1+2	19.5%	43.5%	5.78 (p = 0.0162)
Total	1	16.5%	37.5%	1.81 (n.s.)
Total	2	22.9%	46.7%	3.27 (p = 0.0706)
Dutch	1+2	22.5%	66.7%	8.97 (p = 0.0027)
Dutch	1	23.3%	73.1%	4.29 (p = 0.0383)
Dutch	2	21.6%	62.5%	4.89 (p = 0.0271)

Notes: $P(CPT=1 \mid MUI=0)$ denotes the probability that a respondents satisfies constant proportional trade-off given that he/she does not satisfy *MUI*. The model χ^2 has been calculated by the Likelihood Ratio test.

Table 4.5 displays the results of the estimation procedure. Estimation results have only been reported for the situation where no adjustment for imprecision error was made. Information on whether or not a respondents satisfies utility independence does not improve the model significantly if imprecision adjustment is applied. This

is caused by the unequal distribution of observations over cells,[12] which in turn is a consequence of the fact that with imprecision adjustment a majority of respondents satisfy both constant proportional trade-off and utility independence.

Table 4.5 shows that in every situation respondents who satisfy utility independence are more likely to satisfy constant proportional trade-off. The contribution of the model is significant in all but one case. The pattern is even more clear for the Dutch sample. Here the contribution of the model is always significant. Results for the Swedish sample have not been reported. The reason is that in the Swedish sample information on whether or not a respondent satisfies utility independence did not contribute significantly to the model, even though respondents satisfying utility independence were always more likely to satisfy constant proportional trade-off as well

4.5.2 Group analysis

Constant proportional trade-off

Table 4.6 shows the mean values for the time trade-off questions. Within samples, constant proportional trade-off predicts equality between *D10* and *D30* in version 1 and between *B10* and *B30* in version 2. Table 4.6 shows that the time trade-off weight is slightly higher when 10 years is used as the time horizon. This suggests an increasing proportional trade-off. However the difference is statistically insignificant for both health states.

It is hard to conclude anything definitive about anchoring from the results reported in table 4.6. In the individual analysis we concluded that increasing proportional trade-off and decreasing proportional trade-off are observed with approximately equal frequency. The group analysis indicates a slight tendency to increasing proportional trade-off, but differences are not statistically significant. A slight tendency to increasing proportional trade-off should, if anchoring indeed affects responses, lead to a slightly higher time trade-off weight for *D30* in version 1 and to a slightly higher time trade-off weight for *B30* in version 2. The weight for *D30* is indeed higher in version 1, although the difference is not significant. For version 2 no indications of anchoring exist. The time trade-off weight for *B30* is in fact lower in version 2.

---

[12] By a cell we mean a particular combination of constant proportional trade-off and utility independence, e.g. (CPT fulfilled, UI not fulfilled).

Table 4.6: Mean time trade-off weights (standard error)

Version	Health profile	Total sample (n=172)	Swedish sample (n=80)	Dutch sample (n=92)
1	D 10 yrs.	0.5901	0.5875	0.5924
		(.0194)	(.0270)	(.0280)
1	D 30 yrs.	0.5893	0.5871	0.5913
		(.0201)	(.0269)	(.0296)
1	B 30 yrs.	0.8045	0.7875	0.8190
		(.0155)	(.0243)	(.0198)
2	B 10 yrs.	0.7947	0.7663	0.8201
		(.0169)	(.0237)	(.0235)
2	B 30 yrs.	0.7841	0.7638	0.8022
		(.0179)	(.0261)	(.0246)
2	D 30 yrs.	0.5664	0.5279	0.6006
		(.0273)	(.0389)	(.0378)

The between samples test, which is not susceptible to anchoring bias, also provides support for constant proportional trade-off. We compared the version 1 responses to *D10* with the version 2 responses to *D30* and the version 1 responses to *B30* with the version 2 responses to *B10*. For health state *D* the time trade-off weight for 10 years is slightly higher than for 30 years. For health state *B* the value for 30 years is slightly higher. However, both differences are not significant.

Table 4.6 does not give us reason to suspect that a bias has been introduced by the fact that different health states were included in the two versions. Recall from the argument outlined in the methods section that if this bias affects the results we expect the weights for both *B30* and *D30* to be higher in version 1. Moreover, we expect the difference to be more pronounced for *B30*. Table 4.6 shows that the weights for both *B30* and *D30* are higher in version 1. However, the difference is not significant. Moreover, the difference is not more pronounced for *B30*.

Table 4.6 also displays the results for the Swedish and Dutch samples separately. In both samples the results support constant proportional trade-off: none of the differences is significant. In the Dutch sample higher weights were elicited. However, the difference between the Swedish and the Dutch sample is in no case statistically significant. In the Dutch sample there is no clearcut pattern that indicates problems of anchoring. However, the results of the Swedish sample point

at possible anchoring problems for version 1. The pattern of responses does not raise concerns as to the possible bias that may have been introduced by the fact that different health states were included in the two versions.

Utility independence

Table 4.7 displays the standard gamble weights. The results of the within subjects analysis do not provide support for utility independence. Utility independence predicts equality between *D10* and *D30* in version 1 and between *B10* and *B30* in version 2. However, in both versions the standard gamble weight for 10 years is higher than the standard gamble weight for 30 years. For health state *D* the difference is significant at the 5 % level ($p = 0.0385$), for health state *B* the difference is significant at the 1% level ($p = 0.000$).

We cannot completely dismiss the suspicion that there may have been some anchoring in version 1. We have observed, both in the analysis of the individual responses and in the analysis of the grouped responses, that the standard gamble weight tends to decline the longer the time horizon the assessment is carried out for: the weights for *B30* and *D30* are lower than the weights for *B10* and *D10* respectively. However, anchoring will tend to mitigate the decrease in the quality weight for the 30 years assessment. Thus, given that *D30* is lower than *D10*, anchoring will cause *D30* to be lower in version 2. Table 4.7 shows that this is indeed the case, although the difference is not significant. For version 2 anchoring does not seem to have affected the results. *B30* is in fact slightly lower in version 2 than in version 1.

The between subjects test of utility independence also suggests violation of utility independence. Both *D10* and *B10* are higher than their relative counterparts, which confirms the pattern observed within samples. However, the differences are not significant.[13] This may be due to the lower power of the independent samples t-test.

[13] This may appear somewhat surprising because D30 in version 2 is lower than D30 in version 1 for example and their standard errors are approximately equal. However, within versions the paired t-test is used in which correlation between D10 and D30 is taken into account. Between versions independence of valuations is assumed. The independence assumption results in larger standard errors and therefore lower t-values.

There is no indication that a bias has been introduced by the difference in included health states in the two versions. The weights for *B30* and *D30* are higher in version 2, but the difference is not significant and, contrary to expectation, the difference is more pronounced for health state *D* than for health state *B*.

Table 4.7: Mean standard gamble weights (standard error)

Version	Health profile	Total sample (n=172)	Swedish sample (n=80)	Dutch sample (n=92)
1	D 10 yrs.	0.7017	0.7368	0.6718
		(.0249)	(.0314)	(.0374)
1	D 30 yrs.	0.6784	0.6923	0.6667
		(.0256)	(.0326)	(.0386)
1	B 30 yrs.	0.8651	0.8906	0.8434
		(.0181)	(.0243)	(.0276)
2	B 10 yrs.	0.8972	0.8956	0.8987
		(.0145)	(.0213)	(.0201)
2	B 30 yrs.	0.8507	0.8456	0.8552
		(.0191)	(.0282)	(.0263)
2	D 30 yrs.	0.6597	0.6318	0.6846
		(.0282)	(.0389)	(.0403)

The results for the Swedish and Dutch samples are not significantly different and confirm the above pattern. Within samples, utility independence can be rejected in all but one case (version 1 in the Dutch sample). Between samples, utility independence can only be rejected for health state *D* in the Swedish sample. Some concern exists that the inclusion of different health states has affected the responses in the Swedish sample. Both the weights for *B30* and *D30* are higher (though not statistically significant) for version 1 respondents. The difference is not more pronounced for *B30*, but this could in turn be due to anchoring bias. The biases can explain the rejection of utility independence in the between samples test. Recall from the methods section that the rejection of utility independence in the within samples test cannot be explained by the bias. For the Dutch sample there is no indication that the results have been affected by any of the two biases.

4.6 Discussion

In this chapter we have experimentally tested two of the preference conditions that underly the QALY model in the derivation by Pliskin et al. (1980) to: utility independence and constant proportional trade-off . The results of this chapter suggest that constant proportional trade-off is a condition that describes individual preferences reasonably well. A comparison of mean values, i.e. group analysis, revealed that both within and between subjects constant proportional trade-off could not be rejected. Deviations from constant proportional trade-off are not systematic: increasing proportional trade-off and decreasing proportional trade-off were observed with approximately equal frequency. The analysis of the individual responses showed that 22.8% of the respondents satisfied constant proportional trade-off without adjustment for imprecision error. After adjustment for imprecision error, which is likely to occur given respondents' relative unfamiliarity both with the health states and with the methods of utility measurement, the proportion of respondents whose choices satisfy constant proportional trade-off increased to 90.1%.

Our results provide less support for utility independence. Within subjects utility independence could be rejected at a significance level of 5%. The fraction of the utility of full health decreased rather than to stay constant as predicted by utility independence. The between subjects analysis also suggested violation of utility independence. However, in this case the violations were not statistically significant and we could not reject utility independence. This is probably due to the lower power of the independent samples t-test. The analysis of the individual responses showed that 13.4% of the respondents satisfied utility independence without adjustment for imprecision error. After adjustment for imprecision error this proportion increased to 75.6%

Pliskin et al. (1980) have derived that imposing both constant proportional trade-off and (mutual) utility independence ensures that individual preferences over lotteries over chronic health profiles can be represented by a risk-adjusted QALY model. In our study 5.8% of the respondents satisfy both constant proportional trade-off and utility independence (of quality of life from quantity of life) when no adjustment is made for imprecision error. When adjustment is made for imprecision error this proportion increases to 68.7%.

Adjustment for imprecision error turns out to have an important influence on the results of the individual analysis. It should be reminded that for those respondents who did not indicate a personal confidence interval, a personal confidence interval had to be estimated. Estimation of a personal confidence interval

is necessarily an arbitrary exercise. However, in our opinion it is unlikely that the actual personal confidence intervals are wider than the estimated personal confidence intervals for these respondents. First, the fact that an exact response was given, even though the possibility of indicating personal confidence interval was pointed out to respondents repeatedly, suggests that these respondents had reasonably precise preferences. Second, by the design of our experiment in which the two questions were asked one after the other, the similarity between the two questions was highlighted. Even if the preferences of these respondents are imprecise, it is still probable that in case they considered the two choices to be similar, they express this by giving the same answer. That is, given the design of our experiment we would expect the imprecision error to work in the same direction for these responses. If respondents gave different answers to the two questions, these probably reflect true differences rather than differences due to imprecision error.

We therefore believe that our estimates should be considered as maximum estimates. The partially adjusted results (in which no artificial personal confidence intervals have been constructed and only reported personal confidence intervals have been used) reported in the appendix provide some additional insight. However, as has been emphasized throughout the chapter we find it hard to believe that respondents can give precise answers to the questions that are posed in time trade-off and standard gamble tasks. The fact that a large proportion of respondents did not indicate a personal confidence interval even though we encouraged them to do so is surprising. This may be due to the fact that we did not require respondents to indicate a personal confidence interval, but only included this as an option. Indicating an interval is not necessarily easier than indicating one value. Indication of an interval requires careful thinking about upper and lower bounds. The respondents who did not state an interval may have found the cognitive effort to provide just one value less demanding. It may be necessary in future research to require that respondents indicate a personal confidence interval.

In general we found no indications that our results have been affected by anchoring bias or that a bias arose due to the fact that different versions of the questionnaire included different health states. The only reason for concern are the results of the test of utility independence in the Swedish sample. However, even if we exclude these results and focus for the anlyses that may have been affected by the bias only on the Dutch data, that showed no indications of bias, the above conclusions still hold. This strengthens our belief that the conclusions drawn in this chapter are valid and are not mere artefacts of biases in the experimental design.

The results of this chapter suggest that constant proportional trade-off and utility independence hold approximately and that divergence is due to imprecision

error. These findings have important implications for the use of QALY type utility measures in medical decision making and health policy. Previous findings have indicated that risk neutrality on life years does not describe individual preferences well. Rejection of risk neutrality on life years implies rejection of the risk neutral QALY model. However, our results, in combination with previous research by Miyamoto and Eraker (1988) show that rejection of the risk neutral QALY model does not imply that the whole concept of QALYs has to be dismissed. The support we found for constant proportional trade-off and utility independence suggests that a risk-adjusted QALY model performs well in describing individual preferences for health. The implication of this may be that rather than using a risk neutral QALY model, health researchers should switch to a risk adjusted QALY model.

Appendix 1: The explanation of the time trade-off questions

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

- X years in a specific health state followed by death

- Y years in full health followed by death

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

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

Appendix 2: The explanation of the standard gamble questions

A standard gamble consists of two alternatives:

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

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

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

One way to answer the standard gamble question is to indicate with a - sign those values of p for which you definitely prefer the certain health state and with a + sign

those values of p for which you definitely prefer the treatment option. Finally indicate with a * sign those values of p for which you find it hard to choose between the two profiles.
Next to the line with probabilities of successful treatment a line has been drawn that shows the corresponding probabilities of failure of treatment. This has been done in order to remind you what your choices imply in terms of the probability of failure of treatment.

Appendix 3: Results without the construction of artificial confidence intervals

This appendix displays the results under the assumption that respondents who did not indicate a personal confidence interval were certain of their response. That is, no artificial confidence intervals are constructed.

Table A2: Number of time trade-off responses per category (partial adjustment for imprecision error)

Sample	Version	n	A	B	C	D	E
Total	1	87	27	9	16	8	27
Total	2	85	27	5	23	3	27
Swedish	1	40	16	5	3	3	13
Swedish	2	40	11	3	10	1	15
Dutch	1	47	11	4	13	5	14
Dutch	2	45	16	2	13	2	12

Note: A = weight 10 years > weight 30 years and no overlap PCIs.
B = weight 10 years > weight 30 years, but overlap PCIs
C = weight 10 years = weight 30 years
D = weight 10 years < weight 30 years, but overlap PCIs
E = weight 10 years < weight 30 years and no overlap PCIs

Table A2 shows the categorization of the time trade-off responses. This table is comparable to table 4.2 in the main text. Table A2 shows that 64 respondents (37.2%) satisfy constant proportional trade-off with partial adjustment for imprecision error. In the Dutch sample 39 respondents (42.4%) satisfy constant proportional trade-off with partial adjustment for imprecision error. In the Swedish sample 25 respondents (31.3%) satisfy this condition.

Table A3 shows the categorization of the standard gamble responses. This table is comparable to table 4.3 in the main text. Table A3 shows that 31 respondents (18.0%) satisfy utility independence with partial adjustment for imprecision error. This is only slightly higher than the number of respondents who satisfy utility independence when no adjustment is made for imprecision error. Seventeen Dutch respondents (18.5%) and 14 Swedish respondents (17.5%) satisfy utility independence with partial adjustment for imprecision error.

Table A3: Number of standard gamble responses per category (partial adjustment for imprecision error)

Sample	Version	n	A	B	C	D	E
Total	1	87	49	3	8	2	25
Total	2	85	54	2	15	1	13
Swedish	1	40	26	1	4	0	9
Swedish	2	40	25	2	7	0	6
Dutch	1	47	23	2	4	2	16
Dutch	2	45	29	0	8	1	7

Note: A = weight 10 years > weight 30 years and no overlap PCI's.
B = weight 10 years > weight 30 years, but overlap PCI's
C = weight 10 years = weight 30 years
D = weight 10 years < weight 30 years, but overlap PCI's
E = weight 10 years < weight 30 years and no overlap PCI's

Finally, table A4 shows the number of respondents who satisfy constant proportional trade-off and utility independence with partial adjustment for imprecision error. In the total sample 18 respondents (10.5%) satisfy the two conditions simultaneously. Twelve respondents (13.0%) in the Dutch sample and 6 respondents (7.5%) in the Swedish sample satisfy both conditions.

Table A4: number of respondents satisfying constant proportional trade-off and utility independence simultaneously.
(partial adjustment for imprecision error)

Sample	Version	n	$CPT_{adj}+MUI_{adj}$
Total	1	87	6
Total	2	85	12
Swedish	1	40	1
Swedish	2	40	5
Dutch	1	47	5
Dutch	2	45	7

5

Explaining the disparity between extreme and assorted standard gambles[1]

Summary

Previous research has indicated that extreme standard gambles, in which full health and immediate death are used as gamble outcomes, lead to different health state utilities than assorted gambles in which immediate death is replaced as the worst gamble outcome. The objective of this study was to examine by means of an experiment three explanations for this observed disparity: (i) a framing effect: extreme and assorted gambles invoke different evaluation strategies; (ii) imprecision of preferences: because of the unfamiliarity of the task and of the health states to be evaluated, respondents' preferences are likely to be imprecise. Not taking this imprecision into account may lead to the observed disparity; (iii) probability weighting: respondents do not evaluate probabilities linearly, but transform probabilities to decision weights. The results showed that a correction for the hypothesized framing effect could not explain the observed disparity. A majority of respondents did have imprecise preferences. However, even after allowing for imprecision of preference the disparity remained systematic. Adjustment for probability weighting removed the systematic disparity. A pessimistic weighting probability function, reflecting a high degree of risk aversion, by which favourable outcomes receive a relatively low weight was most consistent with the experimental data. The probability weighting function proposed by Tversky and Kahneman performed worse than not weighting probabilities, which corresponds to expected utility theory. The bad performance of the weighting function proposed by Tversky and Kahneman in this study can be explained by the fact that probability equivalence methods, such as the standard gamble as it is used in health state valuation, lead to extreme risk averse behaviour. An implication of extreme risk averse behaviour is that the standard gamble will assign utilities to health states that are too concave.

[1] Based on Bleichrodt, H., "Explaining the disparity between extreme and assorted standard gambles" (submitted for publication).

5.1 Introduction

During the past two decades, several researchers interested in estimating utilities for health states have reported evidence that the methods commonly employed to measure these utilities give different results.[2] Given common scaling (generally the utility of full health is set equal to one and the utility of immediate death is set equal to zero) the typical pattern is that utilities elicited by the standard gamble (SG) are relatively higher than utilities elicited by the time trade-off (TTO), which in turn are relatively higher than utilities elicited by the rating scale (RS). That is, the standard gamble leads to utilities that are relatively more concave than the utilities elicited by the time trade-off, which are relatively more concave than the utilities elicited by the rating scale. The disparities between the methods for utility elicitation may have the worrying implication for cost utility analysis that different policies are recommended depending on which method is used.

Given that different methods produce different results, the question arises which method should be preferred. Several authors have suggested that the standard gamble should be regarded as a criterion method [e.g. Weinstein, Fineberg et al., 1980; Torrance and Feeny, 1989]. This view is based on the fact that the standard gamble has a well-established axiomatic foundation, being rooted in von Neumann Morgenstern expected utility theory (*EU*) [von Neumann and Morgenstern, 1944]. The assessment of the standard gamble as the criterion method in measuring health state utilities depends crucially on the validity of the axioms underlying von Neumann Morgenstern expected utility theory. Experimental evidence involving monetary outcomes has shown that individuals systematically violate these axioms [Kahneman and Tversky, 1979; Schoemaker, 1982; Harless and Camerer, 1994]. The results of these studies challenge expected utility theory as a descriptive theory of decision under risk.

In valuing health states, the standard gamble method requires an individual to compare two options: a health state for certain and a treatment option with two possible outcomes, one corresponding to successful treatment and one corresponding to failure of treatment. The individual varies the probability of successful treatment until indifference holds between the options. The utility of the health state for certain is then by expected utility theory derived from this probability. If an individual's preferences satisfy the von Neumann Morgenstern axioms, the utility assigned to a health state by the standard gamble should be independent of the health

[2] Cf. e.g. Torrance, 1976; Quinn, 1981; Wolfson et al., 1982; Read et al., 1984; Llewellyn-Thomas et al., 1984; Hershey and Schoemaker, 1985;Hornberger et al., 1992; Stiggelbout et al., 1994; Rutten-van Mölken et al., 1995; chapter 6 of this thesis.

states used in the treatment option [Farquhar, 1984]. Suppose an individual is indifferent between health state X for certain and a treatment option offering a probability of 0.6 to restore his health to full health and a probability of 0.4 of immediate death. Because the von Neumann Morgenstern utility function is unique up to positive affine transformations, the utility function can be scaled such that the utility of full health is equal to 1 and the utility of immediate death is equal to 0. The (expected) utility of X is then equal to

$$U(X) = 0.6*U(full\ health) + 0.4*U(immediate\ death) = 0.6$$

Suppose further that in a comparison between health state Y for certain and the same treatment option giving full health and immediate death as possible outcomes, the individual has indicated to be indifferent between these two options for a probability of restoring his health to full health of 0.2. That is, $U(Y)$ is equal to 0.2. According to expected utility theory, if the utility of health state X is assessed by replacing immediate death by health state Y in the treatment option, then the individual should adjust his indifference probability such that $EU(X)$ is still equal to 0.6. That is, if

$$0.6 = p*1 + (1-p)*0.2$$

then, in order to be consistent with expected utility theory, the individual should state an indifference probability of 0.5.

Following Farquhar (1984), standard gambles in which full health and immediate death are the reference health states will be referred to as extreme gambles, as opposed to assorted gambles in which full health and/or immediate death has been replaced by intermediate health states. Llewellyn-Thomas et al. (1982) observed that, given common scaling, assorted gambles in which immediate death had been replaced as the outcome of unsuccessful treatment resulted in significantly higher health state utilities than extreme gambles. These results challenge expected utility theory as a descriptive theory of health decision making under risk.

The aim of this chapter is to consider three, not necessarily mutually exclusive, explanations for why disparities between extreme and assorted gambles may have been observed: a framing effect, imprecision of preferences, and probability weighting. Section 2 describes these explanations in more detail. Section 3 describes an experiment aimed at testing the three explanations. In section 4 the methods used to analyze the experimental data are described. Section 5 discusses the results of the experiment. Section 6 contains concluding remarks.

5.2 Explanations

The first explanation for the observed disparity between utilities elicited by extreme gambles and by assorted gambles concerns the *potential bias* introduced by the framing of these gambles. This explanation is suggested by the research findings of Llewellyn-Thomas et al. (1982). Llewellyn-Thomas et al. only observed the disparity for assorted gambles in which the worst outcome of treatment changed. When the best outcome of treatment changed the disparity was not observed and the utilities elicited by extreme and assorted gambles did not differ significantly. This suggests that the disparity is caused by a difference in evaluation between extreme gambles and assorted gambles in which the worst outcome of treatment changes. In extreme gambles the common procedure is to ask respondents to give an indifference probability of successful treatment. This may lead respondents to focus on the best outcome of treatment. On the other hand, in assorted gambles where the worst outcome varies over gambles, respondents may be inclined to focus on the distinguishing feature of successive gambles, i.e. the worst outcome of treatment. Taking the certain health state as a reference point, the best outcome of treatment is seen as a gain in health, but the worst outcome of treatment is seen as a loss in health. It is well known from the psychological literature that losses loom larger than gains (loss aversion) [e.g. Kahneman and Tversky, 1979]. The change of focus from the best outcome of treatment to the worst outcome of treatment may thus lead to greater risk aversion. This greater risk aversion is translated in higher indifference values for p, leading to higher utilities for health states.

The second explanation for the observed disparities between extreme and assorted standard gambles is the *imprecision of people's preferences* [Dubourg et al., 1994]. In answering standard gamble questions respondents are asked to perform a task they are not familiar with: trading off risk against improvements in health . Further, respondents have to answer the standard gamble questions within a limited time period without much opportunity to consider the questions thoroughly. Finally, respondents are frequently asked to imagine health states which they have not actually experienced. Under these circumstances it should not come as a surprise that respondents are not able to give a precise probability for which indifference holds, but are only able to identify an interval within which this probability lies. The imprecision of preference hypothesis asserts that in answering an extreme standard gamble question, a respondent selects a probability from an interval, which can be considered as his/her personal confidence interval. We denote the lower and upper bounds of this personal confidence interval by E_L and E_U respectively. When asked an assorted gamble question the respondent selects a probability from the interval with lower and upper bounds A_L and A_U respectively. It might well be that the

personal confidence intervals for the extreme and assorted gambles do not exactly coincide, for example because of the presence of some framing bias. However, under expected utility theory one would expect that the personal confidence intervals do overlap.

The third explanation is *probability weighting.* Probability weighting is a distinguishing feature of rank dependent utility theory (*RDU*), first introduced by Quiggin (1982) and presently the most popular alternative to expected utility theory as a theory of decision under risk. In expected utility theory probabilities are entered linearly, i.e. a gamble is evaluated by $\sum_{j=1}^{n} p_j U(x_j)$ where p_j denotes the probability that event j occurs and x_j denotes the resulting outcome when event j occurs. In rank dependent utility theory probabilities p are transformed by a weighting function $\pi(p)$. The transformed probabilities share with probabilities the properties that $\pi(0)=0$; $\pi(1)=1$; and if $p_1 > p_2$ then $\pi(p_1)>\pi(p_2)$. However, transformed probabilities differ from probabilities in that they are in general not additive, i.e. $\pi(p_1+p_2)\neq\pi(p_1)+\pi(p_2)$. Notice that expected utility theory corresponds with the case where the weighting function is the identity function, i.e. $\pi(p) = p$.

The rank ordering of the outcomes is crucial in calculating the rank dependent utility of a gamble. Suppose the outcomes of a gamble x are rank ordered in increasing order of preference, i.e. $x_1 \leq x_2 \leq \ldots. \leq x_n$ in the case of n different outcomes, where $\leq$ stands for "not strictly preferred to." Then the rank dependent utility of this gamble is

$$\sum_{j=1}^{n} \pi_j U(x_j) \qquad (1)$$

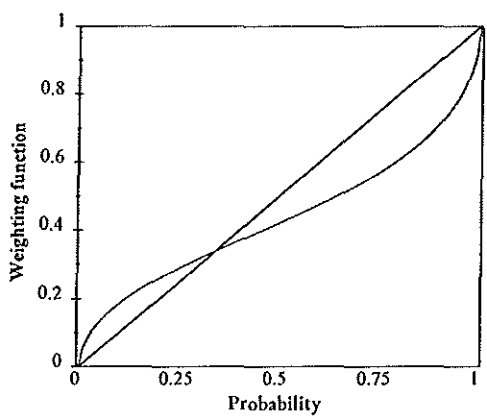
where the decision weights π_j are defined as: $\pi_j = \pi(\sum_{i=j}^{n} p_i) - \pi(\sum_{i=j+1}^{n} p_i)$. As an example consider an extreme gamble giving full health with probability p and immediate death with probability $(1-p)$. The rank ordering of the outcomes of this gamble is: *immediate death* $\leq$ *full health*. Because the utility function in rank dependent utility theory is unique up to positive affine transformations, we can apply the scaling U *(full health)* $= 1$ and U *(immediate death)* $= 0$. The rank dependent utility of this extreme gamble can then, by application of the above formula, be calculated as: $\pi(p)*U(\textit{full health}) + [\pi\{p+(1-p)\}-\pi(p)]*U(\textit{immediate death}) = \pi(p)*1+(1-\pi(p))*0=\pi(p)*1$. Wakker and Stiggelbout (1995) provide a more extensive discussion of the application of rank dependent utility theory in medical decision making. Rank dependent utility theory is equivalent to cumulative prospect theory when all outcomes are gains [Tversky and Kahneman, 1992].

Several authors have suggested, on the basis of empirical research findings for monetary outcomes, that the weighting function $\pi(p)$ is S-shaped.[3] More specifically, Tversky and Kahneman (1992) have suggested that the following weighting function is plausible for the "average individual" when all outcomes concern gains:

$$\pi(p) = \frac{p^{.61}}{\left[p^{.61} + (1-p)^{.61}\right]^{1/.61}} \qquad (2)$$

This weighting function has been displayed graphically in figure 5.1. The weighting function displayed in figure 5.1 lies above the diagonal for small probabilities and under the diagonal for larger probabilities. The S-shaped weighting function overweights small probabilities and underweights high probabilities. In the middle range the function is approximately linear, which is consistent with expected utility theory.

Figure 5.1: The S-shaped weighting function proposed by Tversky and Kahneman



Let us now return to the example given in the introduction. Suppose the S-shaped weighting function hypothesized by Tversky and Kahneman holds. The individual has given two indifference probabilities. In a comparison between health state X for certain and an extreme gamble, the individual has stated an indifference probability p of 0.6. In a comparison between health state Y for certain and an extreme gamble the

[3] Cf. e.g. Quiggin, 1982; Karni and Safra, 1990; Tversky and Kahneman, 1992; Camerer and Ho, 1994; Wakker and Stiggelbout, 1995; Tversky and Wakker, 1995.

individual has stated an indifference probability p of 0.2. The utility function is scaled such that U *(full health)* = *1* and U *(immediate death)* = *0*. The expected utility of health state X is then equal to p, thus $EU(X) = 0.6$. The rank dependent utility of health state X, applying the formula for rank dependent utility theory given above, is equal to $\pi(0.6)*1+(1-\pi(0.6))*0 \approx 0.47$. The expected utility of health state Y is equal to 0.2. The rank dependent utility of health state Y is equal to $\pi(0.2)*1+(1-\pi(0.2))*0 \approx 0.26$. The rank dependent utility of health state Y is higher than the expected utility of health state Y and the rank dependent utility of health state X is lower than the expected utility of health state X. This corresponds to the overweighting of low probabilities and the underweighting of higher probabilities.

In the assorted gamble immediate death is replaced by health state Y as the outcome of unsuccessful treatment. By rank dependent utility theory, the rank dependent utility of health state X should be equal to the rank dependent utility of the assorted gamble giving full health with probability p and health state Y with probability *1-p*. That is, $0.47=\pi(p)*1+(1-\pi(p))*0.26$ or $\pi(p) \approx 0.284$. Thus $p=\pi^{-1}(0.284) \approx 0.24$. Rank dependent utility theory with the probability weighting function suggested by Tversky and Kahneman predicts that this individual will state an indifference probability of 0.24. Suppose the respondent does indeed state an indifference probability of 0.24 in the above assorted gamble. Then the expected utility of health state X is calculated as $0.24*1+0.76*0.20=0.39$. This is clearly different from 0.6, the expected utility of health state X calculated from the extreme gamble. The above example shows that probability weighting can explain why extreme and assorted gambles lead to different expected utilities even though in the above example the pattern is not consistent with the research findings by Llewellyn-Thomas et al. that assorted gambles lead to higher expected utilities than extreme gambles.

5.3 Experimental design

This section describes an experiment aimed at testing to what extent the three explanations, outlined in the above section, can contribute to a better understanding of the observed disparity between the utilities elicited by extreme and assorted gambles. The respondents participating in the experiment were 66 students following a four-year course in health policy and management at the Erasmus University Rotterdam. Respondents were confronted with 5 health state decriptions plus full health and immediate death. The five health state descriptions were selected from the set of health states used by the EuroQol group (1990) and are described in the

appendix. Respondents were asked to imagine the health states to last for the rest of their lives.

Respondents were first asked to rank the health states and subsequently to put the health states on a rating scale calibrated between full health and immediate death. The principal intention of this task was to familiarize the respondents with the health states. The standard gamble method was subsequently explained and respondents were taken through three monetary gambles in order to improve their understanding of the method.

To test the first explanation given in section 2, that the observed disparity is caused by a framing bias, an attempt was made to control for its possible impact. Respondents were confronted not only with a line with probabilities of successful treatment, but also with a line with probabilities of unsuccessful treatment. If the explanation given in section 2 is correct, it was hoped that this design would shift the focus of respondents such that they take into account both the outcome of successful treatment and the outcome of unsuccessful treatment. Under the particular framing bias hypothesized in section 2, this should reduce the disparities between extreme and assorted gambles.

To test the second explanation, that the disparities are caused by imprecision in preferences, respondents were asked to indicate first the probabilities for which they definitely preferred the certain health outcome, then to indicate the probabilities for which they definitely preferred the treatment option, and finally the probabilities for which they could not choose between the two options. It was pointed out repeatedly to respondents that they were allowed to give a range of indifference values for p and that imprecise preferences were thus allowed.

Table 5.1: The 13 SG questions

Number gamble	Certain outcome	Gamble outcomes	Number gamble	Certain outcome	Gamble outcomes
1.	*A*	*F and D*	8.	*D*	*F and C*
2.	*B*	*F and D*	9.	*E*	*F and C*
3.	*C*	*F and D*	10.	*G*	*F and C*
4.	*E*	*F and D*	11.	*B*	*F and G*
5.	*G*	*F and D*	12.	*E*	*F and B*
6.	*A*	*F and C*	13.	*G*	*F and A*
7.	*B*	*F and C*			

Respondents answered 13 standard gamble questions in total, both extreme and assorted. The 13 standard gamble questions are displayed in table 5.1. The questions with health state *C* as the outcome of unsuccessful treatment were included to examine whether the disparity between extreme and assorted gambles could also be observed when death was replaced by a health state which is about as bad as death. The decision not to vary the best outcome of treatment was taken on the basis of previous research findings [Llewellyn-Thomas et al., 1982]. It was judged that the possible gain of additional information did not outweigh the risk of respondents getting bored and becoming less careful in answering the questions. The order of the standard gamble questions varied across respondents. The intention was to avoid possible effects of the order in which the questions were asked.

5.4 Methods

Assorted standard gamble utilities were calculated using the extreme standard gamble utilities as the utilities of the outcomes of the treatment option. For example, in the calulation of the assorted utility for health state *B* by the formula $p*1 + (1-p)*U(G)$, $U(G)$ was set equal to the utility for health state *G* elicited by the extreme gamble. Responses were analyzed both at the aggregate level, i.e. by comparing mean values, and at the individual level. Respondents' personal confidence intervals were treated as ordinary statistical confidence intervals. That is, in calculating mean values we used the midpoint of the personal confidence interval. Statistical significance of the differences in mean values between the extreme and assorted standard gamble valuations was examined both by means of a paired t-test and by means of a non-parametric, distribution-free technique: Wilcoxon's matched pairs signed ranks test. Even though the paired t-test assumes normality of the stochastic variables, it is fairly robust as long as the variances are equal. We tested for equality of variances by means of an F-test. Because it turned out that equality of variance had to be rejected for health state *E*, we also examined the differences by the non-parametric method.

At the individual level we classified responses in eight different categories (see for example table 5.3). The first three categories consist of those respondents for whom the utility elicited by the extreme standard gamble is smaller than the utility elicited by the assorted standard gamble. Category 1 consists of those respondents who did not indicate a personal confidence interval, but gave exact indifference probabilities. Category 2 consists of those respondents who did indicate a personal confidence interval, but the upper bound of the personal confidence interval for the utility elicited by the extreme standard gamble is lower than the lower bound of the personal confidence interval for the utility elicited by the assorted standard gamble.

That is, the personal confidence intervals do not overlap. Category 3 consists of those respondents for whom the upper bound of the personal confidence interval for the utility elicited by the extreme standard gamble is higher than the lower bound of the personal confidence interval for the utility elicited by the assorted standard gamble. That is, the personal confidence intervals do overlap. Categories 4 and 5 consist of those respondents for whom the utility elicited by the extreme gamble is equal to the utility elicited by the assorted gamble. Respondents in category 4 did not indicate a personal confidence interval. Respondents in category 5 did indicate a personal confidence interval. Respondents in categories 6, 7 and 8 indicated higher utilities by the extreme standard gamble than by the assorted standard gamble. Category 6 corresponds to category 3 in that personal confidence intervals overlap. Category 7 corresponds to category 2 in that personal confidence intervals do not overlap. Category 8 corresponds to category 1: no personal confidence interval is indicated. Categories 3 and 6 contain those respondents who indicate a different utility by the extreme gamble and by the assorted gamble, but for whom this difference is not significant in terms of their personal confidence interval. The observed difference is due to imprecision error for these respondents.

Statistical significance of differences between proportions was examined by a χ^2 test. A continuity correction was applied to take into account that the discrete binomial distribution is approximated by the continuous normal distribution. It is well known that this approximation is better when a correction of ½ is made to the observed frequency [Altman, 1991].

The methods applied are similar for expected utilities and for rank dependent utilities. The difference between the analysis for the expected utilities and the analysis for the rank dependent utilities is that in the former the indicated probabilities enter linearly in the evaluation formula, whereas in the latter the cumulative probabilities are transformed into decision weights.

To illustrate the procedure consider table 5.2 which shows the responses of a particular respondent included in the experiment. The table shows that in the extreme gamble with *B* for certain this respondent has indicated that for probability values lying between 0.75 and 0.80 she had no clear preference for one of the two options. Given the scaling *U (full health) = 1* and *U (immediate death) = 0*, this translates in a personal confidence interval for the expected utility of health state *B* ranging from 0.75 to 0.80. In calculating mean values we set the expected utility of *B* equal to the midpoint of this interval, i.e. *U (B) = 0.775*. In a similar way we determine that *U(G)* lies between 0.45 and 0.50 for this respondent and use *U (G) = 0.475* in the aggregate analysis. In the assorted gamble for health state *B* with *G* as the worst outcome in the treatment option, the respondent has indicated that she did not have a clear preference for one of the two options for probabilities lying between 0.65 and 0.70.

This translates in a lower bound of the personal confidence interval for the expected utility of health state B of $0.65*1+0.35*0.45 = 0.8075$ and an upper bound of $0.70*1+0.30*0.50=0.85$. The midpoint of this interval is 0.8288 which was used in the calculation of the mean values. Given that $0.775 < 0.8288$ this individual falls into one of the first three categories in the individual analysis of the expected utilities. The upper bound of the personal confidence interval for the expected utility elicited by the extreme gamble is 0.80 which is smaller than 0.8075 the latter being the lower bound of the expected utility elicited by the assorted gamble. By consequence this respondent is classified in category 2. We conclude that for this respondent the difference between the expected utility elicited by the extreme gamble and the expected utility elicited by the assorted gamble is systematic and cannot be explained by imprecision error.

Table 5.2: Responses of one particular respondent

	Extreme gamble B	Extreme gamble G	Assorted gamble B (G worst outcome)
p	0.75-0.80	0.45-0.50	0.65-0.70
EU	$0.75 \leq EU(B) \leq 0.80$	$0.45 \leq EU(G) \leq 0.50$	$0.8075 \leq EU(B) \leq 0.850$
RDU	$0.568 \leq RDU(B) \leq 0.607$	$0.395 \leq RDU(G) \leq 0.421$	$0.699 \leq RDU(B) \leq 0.730$

Table 5.2 also shows the rank dependent utilities for the weighting function proposed by Tversky and Kahneman (1992). For example the lower bound of the personal confidence interval of the rank dependent utility of health state B elicited in the extreme gamble is $\pi(p)*1+[1-\pi(p)]*0=\pi(p)= 0.75^{0.61}/\{[0.75^{0.61} + 0.25^{0.61}]^{1/0.61}\} = 0.568$. Given that $0.699 > 0.607$ this respondent is classified in category 2 on the basis of rank dependent utility theory also. The difference between the rank dependent utility elicited by the extreme gamble and the rank dependent utility elicited by the assorted gamble is systematic and cannot be explained by imprecision error.

The impact of order effects was examined both by means of analysis of variance and by means of the distribution-free Kruskal Wallis test.

5.5 Results and discussion.

The answers of 62 respondents were included in the analyses. Four respondents were excluded because they failed to answer one or more of the standard gamble questions.

Both parametric (paired t-test) and non-parametric (Kruskal-Wallis) tests showed no evidence of order effects on the results.

5.5.1 Framing

Table 5.3[4] displays the extreme standard gamble utilities, the assorted standard gamble utilities, their difference and the probability that this difference would be observed under the null hypothesis that the true difference is equal to zero. Table 5.3 shows that the differences between the extreme standard gamble utilities and the assorted standard gamble utilities are significant at the 1% level for health states *B* and *G*. For health state *E* the difference is statistically significant at the 5% level. The reduced statistical significance of the difference for health state *E* could be expected given that the utility elicited by the extreme standard gamble is already close to one for health state *E*. Llewellyn-Thomas et al. also found that the disparity between the utilities elicited by extreme and assorted standard gambles is most pronounced for health states with relatively low utilities.

Table 5.3: Extreme and assorted standard gamble responses

Health state	Utility extreme SG (standard error)	Utility assorted SG (standard error)	Difference	Statistical significance T-test	Wilcoxon
B	0.7798	0.8389	-0.0591	p=0.001	p=0.000
	(0.0200)	(0.0198)	(0.0175)		
E	0.9389	0.9590	-0.0201	p=0.033	p=0.041
	(0.0130)	(0.0086)	(0.0092)		
G	0.6774	0.7500	-0.0726	p=0.002	p=0.001
	(0.0284)	(0.0292)	(0.0229)		

[4] The results presented here are for full health and immediate death as outcomes in the extreme gambles. The results are similar using full health and health state C in the extreme gambles. These results can be obtained from the author on request.

The summary statistics displayed in table 5.3 indicate that the observed disparity between utilities elicited by extreme and assorted gambles is persistent in spite of the attempt made to control for the framing bias hypothesized in section 2. Either the attempt to control for this type of framing bias failed or this framing bias is not important in the explanation of the observed disparity.

5.5.2 Imprecision of preferences

Table 5.4 displays the results of the individual analysis of responses. From this table insight can be obtained to what extent the observed disparities are systematic or are caused by imprecision error. A majority of the respondents indicated a personal confidence interval for health states *B* and *G*, confirming that their preferences are indeed imprecise. This is also true for health state *E* when the 18 respondents who were not willing to take any risk are ignored. The contribution of imprecise preferences to the explanation of the observed disparity between utilities elicited by extreme and by assorted gambles can be assessed by calculating the proportion of respondents in categories 3 and 6. For health states *B, E* and *G* these proportions are 22.6 %, 9.7 % and 21.0 % respectively.

We expect that under expected utility theory with imprecise preferences responses fall in categories 3 to 6. Expected utility theory predicts equality of utilities elicited by extreme and assorted gambles. The imprecise preferences hypothesis allows some disparities between these utilities but if individuals do indeed behave according to expected utility theory, one would expect the personal confidence intervals to overlap. The proportions of respondents satisfying the combination of expected utility theory and imprecise preferences are 25.8 %, 40.3 % and 22.6 % for health states *B, E* and *G* respectively. The proportion of respondents satisfying expected utility theory corrected for imprecise preferences is somewhat distorted for health state E given that 18 respondents were not willing to take any risk. Excluding these respondents leaves only 11.3 % of the respondents satisfying the combination of expected utility theory and imprecise preferences for health state *E*.

Table 5.4 further shows that the disparity between the utilities elicited by extreme standard gambles and by assorted standard gambles is systematic. Categories 1 and 2 contain those respondents for whom the utility elicited by the assorted gamble is significantly higher in terms of their personal confidence interval than the utility elicited by the extreme gamble. Categories 7 and 8 contain those respondents for whom the utility elicity by the extreme gamble is significantly higher in terms of their personal confidence interval than the utility elicited by the assorted gamble. Table 5.4 displays that the proportion of respondents in categories 1 and 2 is higher

than the proportion of respondents in categories 7 and 8. For health states *B* and *G*, the difference between the proportions is significant at the 1 % level. For health state *E* the difference between the proportions is significant at the 10 % level. The pattern which was observed at the aggregate level is thus confirmed at the individual level: there is a systematic disparity between utilities elicited by extreme and assorted gambles. The proportion of respondents who give a significantly higher utility to health states by assorted gambles is higher than the proportion of individuals who give a significantly higher utility to health states by extreme gambles.

Table 5.4: Categories of extreme and assorted standard gamble responses

Category	Health state *B*	Health state *E*	Health state *G*
1. Exact value given, extreme < assorted	16	13	16
2. No overlap PCI, extreme < assorted	19	10	17
3. Overlap PCI, extreme < assorted	10	4	11
4. Exact value given, extreme = assorted	1	18	0
5. PCI given, extreme = assorted	1	0	1
6. Overlap PCI, extreme > assorted	3	3	2
7. No overlap PCI, extreme > assorted	7	10	11
8. Exact value given, extreme > assorted	5	4	4
Ratio (1+2)/total	0.565	0.371	0.532
Ratio (7+8)/total	0.194	0.226	0.242
Significance difference (χ^2)	$p < 0.01$	$p < 0.10$	$p < 0.01$

5.5.3 Probability weighting

As outlined above, expected utility theory with imprecise preferences can only explain 25% of the observed responses and cannot explain the systematic disparity between utilities elicited by extreme gambles and utlities elicited by assorted gambles.

This leaves the third explanation: respondents do not evaluate gambles according to expected utility theory, but according to rank dependent utility theory in which probabilities are transformed into decision weights. Table 5.5 shows that on an aggregate level rank dependent utility theory with the weighting function proposed by Tversky and Kahneman (RDU_{TK}) does not remove the systematic disparity between utilities elicited by extreme gambles and utilities elicited by assorted gambles. If anything, RDU_{TK} performs worse than expected utility theory. The differences between the utilities are approximately twice as large as under expected utility theory and the differences are significant at the 1 % level for all health states.

Table 5.5: Extreme and assorted rank dependent utilities. Weighting function Tversky and Kahneman (1992)

Health states	Extreme utility (standard error)	Assorted utility (standard error)	Difference	Statistical significance T-test	Wilcoxon
B	0.6254	0.7465	-0.1211	p=0.000	p=0.000
	(0.0174)	(0.0173)	(0.0148)		
E	0.8511	0.8952	-0.0441	p=0.002	p=0.005
	(0.0198)	(0.0136)	(0.0139)		
G	0.5507	0.6795	-0.1288	p=0.000	p=0.000
	(0.0208)	(0.0227)	(0.0181)		

In table 5.6 the results for the individual analysis are displayed and these confirm the above pattern. The proportions of respondents satisfying RDU_{TK} with imprecise preferences (categories 3 to 6 in table 5.6) are 19.3 %, 35.4 % and 16.1 % for health states *B, E* and *G* respectively. Excluding the 18 respondents who were not willing to take any risk at all in gambles involving health state *E*, leaves 9.1 % of respondents who satisfy RDU_{TK} with imprecise preferences for health state *E*. Without exception, the proportions of respondents who satisfy RDU_{TK} with imprecise preferences is lower than the proportions of respondents who satisfy expected utility theory with imprecise preferences.

The disparity between the utilities elicited by extreme and assorted gambles is also more systematic for RDU_{TK} than for expected utility theory. Compared to table 5.4 which displays the results for expected utility theory, in table 5.6 the proportion of respondents in categories 1 and 2 is higher for all health states and the proportion of respondents in categories 7 and 8 is lower. The difference between the proportion

of respondents in categories 1 and 2 and the proportion of respondents in categories 7 and 8 is significant at the 1 % level for all health states.

Table 5.6: Categories of extreme and assorted SG responses. Weighting function Tversky and Kahneman (1992)

Category	Health state B	Health state E	Health state G
1. Exact value given, extreme < assorted	19	14	19
2. No overlap PCI, extreme < assorted	26	13	25
3. Overlap PCI, extreme < assorted	8	3	8
4. Exact value given, extreme = assorted	1	18	0
5. PCI given, extreme = assorted	1	0	0
6. Overlap PCI, extreme > assorted	2	1	2
7. No overlap PCI, extreme > assorted	3	10	7
8. Exact value given, extreme > assorted	2	3	1
Ratio (1+2)/total	0.726	0.436	0.710
Ratio (7+8)/total	0.081	0.209	0.129
Significance difference (χ^2)	$p < 0.01$	$p < 0.01$	$p < 0.01$

Given the bad performance of RDU_{TK} in explaining the disparity between the utilities elicited by extreme and assorted standard gambles, we attempted to estimate a weighting function that would fit the data better. The weighting function we estimated was $\pi(p) = p^{\beta}$. For $\beta>0$ this weighting function satisfies $\pi(0)=0$, $\pi(1)=1$ and $\pi(p_1) > \pi(p_2)$ *if* $p_1>p_2$. Thus, the weighting function is a proper weighting function to be used in rank dependent utility theory. Figure 5.2 displays graphically the shape of the weighting function for various values of β. If $\beta=1$ the weighting function is linear in probability and rank dependent utility theory is identical to expected utility theory. For $0<\beta<1$, the weighting function lies everywhere above the diagonal (the weighting function in the figure has bee drawn for $\beta = 0.5$). Such a weighting function corresponds to optimism. As outlined in

section 5.2, in rank dependent utility theory outcomes are ordered in increasing order of preference and decision weights are calculated as $\pi_j = \pi(\sum_{i=j}^{n} p_i) - \pi(\sum_{i=j+1}^{n} p_i)$. Thus in a 50-50 gamble the most preferred outcome receives a weight of $\pi(0.5)$ and the least preferred receives a weight of $1-\pi(0.5)$. If the weighting function lies everywhere above the diagonal $\pi(0.5) > 1-\pi(0.5)$ and thus preferred outcomes are overweighted relative to less preferred outcomes. Therefore such a weighting function is referred to as reflecting optimism. For $\beta > 1$ the weighting function lies everywhere below the diagonal (the weighting function in the figure has bee drawn for $\beta = 2.85$). By a similar argument as outlined above it can be shown that this weighting function underweights preferred outcomes and therefore corresponds to pessimism.

Figure 5.2: Optimistic and pessimistic weighting functions

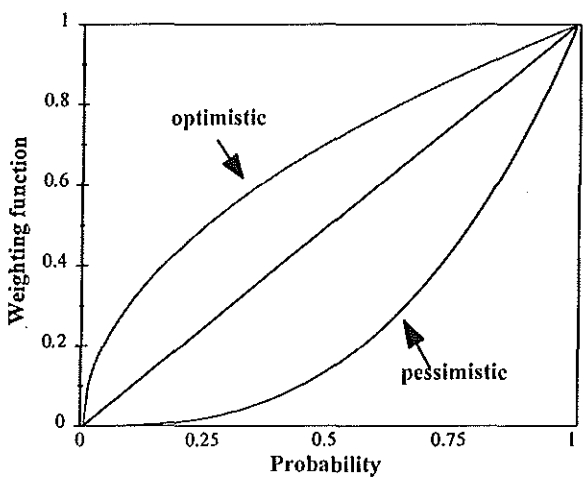


Table 5.7 displays the results of the estimation procedure. Details of the estimation procedure are described in the appendix. The estimated value for β is 2.85 with a standard error of 0.78. This value for β suggests a pessimistic weighting function. The hypothesis that $\beta = 1$, which corresponds to expected utility theory, can be rejected at the 5 % level.

Table 5.7: Estimation results.

	all data	health state *B*	health state *E*	health state *G*
β	2.85	2.50	2.54	4.76
SE(β)	0.78	0.97	1.02	3.37

Table 5.7 also shows estimated weighting functions for the three health states separately. The weighting functions estimated for health states *B* and *E* resemble the overall weighting function. However, for health state *G* the estimated weighting function is quite different. All weighting functions correspond to pessimism. However, the weighting function does not differ significantly from expected utility theory for any of the health states.

Table 5.8: Extreme and assorted rank dependent utilities. Estimated weighting function

Health states	Extreme utility (standard error)	Assorted utility (standard error)	Difference	Statistical significance T-test	Wilcoxon
B	0.5480	0.5678	-0.0198	p=0.0488	p=0.0522
	(0.0311)	(0.0345)	(0.0283)		
E	0.8596	0.8481	0.0115	p=0.0536	p=0.0243
	(0.0245)	(0.0234)	(0.0184)		
G	0.4217	0.4458	-0.0241	p=0.0412	p=0.0418
	(0.0343)	(0.0366)	(0.0291)		

Table 5.8 shows that at the aggregate level, rank dependent utility theory with the weighting function estimated on the basis of the full data set reduces the difference between the utilities elicited by extreme and assorted gambles to insignificance for all health states.

Table 5.9 summarizes the results at the individual level. The proportion of respondents satisfying rank dependent utility theory with the estimated pessimistic weighting function (RDU_P) and allowing for imprecise preferences are 32.2 %, 38.6 % and 25.9 % for health states *B*, *E* and *G* respectively. Thus, allowing for imprecise preferences, RDU_P performs slightly better than expected utility theory for health states *B* and *G* and slightly worse for health state *E*. The disparity between utilities elicited by extreme and assorted standard gambles is no longer systematic. The proportion of respondents in categories 1 and 2 is higher than the proportion of respondents in categories 7 and 8 for health states *B* and *G* and lower for health state

E. The differences between the proportions are not significant for any of the three health states.

Table 5.9: Categories of extreme and assorted SG responses. Estimated weighting function.

Category	Health state *B*	Health state *E*	Health state *G*
1. Exact value given, extreme < assorted	15	9	14
2. No overlap PCI, extreme < assorted	10	7	10
3. Overlap PCI, extreme < assorted	9	3	12
4. Exact value given, extreme = assorted	1	18	0
5. PCI given, extreme = assorted	1	0	0
6. Overlap PCI, extreme > assorted	9	3	4
7. No overlap PCI, extreme > assorted	11	14	16
8. Exact value given, extreme > assorted	6	8	6
Ratio (1+2)/total	0.403	0.258	0.387
Ratio (7+8)/total	0.274	0.355	0.355
Significance difference (χ^2)	n.s.	n.s.	n.s.

5.6 Concluding remarks

The aim of this chapter was to examine three explanations for the observed disparity between utilities elicited by extreme and by assorted standard gambles: a type of framing effect, imprecision of preferences and probability weighting. The main findings of this chapter are:

1. Controlling for a hypothesized framing effect did not reduce the disparity compared to earlier studies. This suggests that the hypothesized framing effect does

not play a pivotal role as a cause of the disparity. However the possibility cannot be excluded that our attempt to control for this framing effect was not successful.

2. The preferences of a majority of respondents were imprecise. This imprecision is hypothesized to be caused by the unfamiliarity of the task and of the health states respondents are confronted with in standard gamble questions. Imprecision of preferences could explain part of the observed disparity. However, the majority of responses could not be explained by expected utility theory and imprecise preferences.
3. Respondents do not enter probabilities linearly in the evaluation of gambles as implied by the axioms of expected utility theory. Instead they transform probabilities. Transformation of probabilities is a distinguishing feature of rank dependent utility theory. However, rank dependent utility theory with the probability weighting function proposed by Tversky and Kahneman (1992) performed worse than expected utility theory both at the aggregate level and at the individual level allowing for imprecision of individual preferences. Rank dependent utility theory with a weighting function corresponding to pessimism (RDU_P), which was estimated on the basis of the experimental data, described individual choices better. Both at the aggregate and at the individual level allowing for imprecise preferences, RDU_P was able to explain the systematic disparity between the utilities elicited by extreme and assorted gambles. Allowing for imprecise preferences, 20-40 percent of individual responses satisfied RDU_P.

The finding that RDU_P fits the data better than RDU_{TK} is surprising given the good performance of the latter in explaining risky choices with respect to monetary outcomes. Camerer and Ho (1994), for example, observe with respect to the elicitations of probability transformations that "the results were remarkably stable across studies." The poor performance of RDU_{TK} most probably indicates that other factors than probability weighting alone cause deviations from expected utility theory and that these factors work in the direction of pessimism.

When we compare the S-shaped probability transformation function with the estimated probability transformation function it appears that the overweighting of small probabilities, implied by the S-shaped function, is absent. Respondents are pessimistic over the whole range of probabilities. Moreover, as can for example be seen from figure 5.2, they are pessimistic to a considerable degree. A reason for this may be that the way standard gamble questions are generally asked for health states, by means of probability equivalence, leads to strong risk averse behaviour. Evidence on this phenomenon is available from the literature [Hershey and Schoemaker, 1985]. In the assorted gambles two probability equivalence questions were implicitly used: one to determine indifference between the certain outcome and the treatment option

and one to determine the utility of the outcome that was substituted for immediate death in the treatment option. Using two probability equivalence questions to determine utilities may lead to extreme risk averse behaviour. Extreme risk averse behaviour is consistent with the estimated weighting function. In short, the estimated weighting function suggests that probability equivalence methods as they are used in health state valuation may give rise to strong risk averse behaviour. Given this, one should be careful to use probability equivalence methods to value health states, because they are likely to lead to utilities that are too high.

Our results do not imply that cumulative prospect theory is rejected for the present study. We cannot infer much about the validity of cumulative prospect theory because we only used the version of cumulative prospect theory that corresponds to a situation where all outcomes are gains.[5] It might be that reference effects influenced the results and this should be taken into account in cumulative prospect theory. Our results showed that rank dependent utility theory, in which no distinction is made for the sign of outcomes, with the weighting function proposed by Tversky and Kahneman for gains did not perform well in explaining the experimental data presented in this chapter.

The disparity between utilities elicited by extreme and assorted gambles spells doubts about the descriptive validity of expected utility theory in health state valuation. RDU_P appears to describe individual preferences better, at least in the context of this study. This does not imply that RDU_P should replace expected utility theory in health state utility measurement. Health state utility measurement should be based on a theory that is normatively valid even though the measurement of utilities must obviously fully reckon with the descriptive deviations from rationality.. The results of this study do not imply anything about the normative validity of expected utility theory. In fact, we asked ten of the respondents several weeks after the experiment to participate in a second interview session. In this session we explained the logic of expected utility theory and asked these respondents to reconsider their earlier choices. Only 3 out of ten respondents were willing to change their earlier responses in order to behave according to *EUT*. Obviously one cannot conclude anything definitive from this finding due to the very small sample size. However, more research into the normative appeal of expected utility theory in health state utility measurement appears to be worthwhile.

The main implication of this chapter is a note of caution. We have obtained indications that asking standard gambles by probability equivalence, as is typically done in health utility measurement, can lead to extreme risk averse behaviour. Extreme risk averse behaviour leads to utilities that are too high. Further, the results

[5] The version corresponding to the situation where all outcomes are losses did not perform better.

of this study do suggest that probability weighting and imprecise preferences have an important impact on health state utilities. Probability weighting implies that one should not enter probabilities elicited from standard gamble questions linearly in expected utility calculations, but that probabilities should first be transformed. Imprecision of preference implies that it appears indispensable, if one is to have confidence in the reliability of responses to standard gamble questions, to allow respondents to become familiar with the idea of trading off probabilities and with the health states they are supposed to evaluate and to give them the opportunity to consider their choices thoroughly. This will make the utility assessment task more involved. However, not taking probability weighting and imprecision of preferences into account is likely to produce unreliable results that may mislead policy decisions.

Appendix 1: Descriptions of the health states

A

1. Unable to walk without a stick, crutch or walking frame
2. No problem with self-care
3. Unable to perform main activity work, study, housework)
4. Unable to pursue family and leisure activities
5. Extreme pain or discomfort
6. Anxious or depressed

B

1. No problems in walking about
2. No problems with self-care
3. Unable to perform main activity work, study, housework)
4. Able to pursue family and leisure activities
5. Extreme pain or discomfort
6. Not anxious or depressed

C

1. Confined to bed
2. Unable to feed self
3. Unable to perform main activity (e.g. work, study, housework)
4. Unable to pursue family and leisure activities
5. Extreme pain or discomfort
6. Anxious or depressed

E

1. No problems in walking about
2. No problems with self-care
3. Able to perform main activity work, study, housework
4. Able to pursue family and leisure activities
5. Moderate pain or discomfort
6. Not anxious or depressed

D

Immediate death

F

Full health

G

1. No problems in walking about
2. No problems with self care
3. Unable to perform main activities work, study, housework)
4. Unable to pursue family and leisure activities
5. Moderate pain or discomfort
6. Anxious or depressed

Appendix 2: Estimation method

The equation to be estimated is

$$p_{1i} = [p_{3i}^{\beta} + (1-p_{3i}^{\beta}).p_{2i}^{\beta}]^{1/\beta} + w_i \qquad (A1)$$

Suppose for example that B is the certain health state to be evaluated in an assorted gamble with full health and health state G as outcomes. Then p_1 is the probability elicited in an extreme gamble with B as the certain health state, p_3 is the probability elicited in the assorted gamble with B as the certain health state and p_2 is the probability elicited in an extreme gamble with G as the certain outcome. $RDU(B) = 1*p_1^{\beta}$, $RDU(G) = 1*p_2^{\beta}$, and by consistency $RDU(B)$ is also equal to $1*(p_3^{\beta} + (1-p_3^{\beta})*p_2^{\beta}$. The equation to be estimated now follows. This equation can be estimated by non-linear least squares. The procedure we used was the Gauss Newton method, which is an iterative procedure [Greene, 1993].

A problem arises if we assume that rank dependent utilities are measured with error. In that case the covariance between p_3 and w is not equal to zero, which is one of the assumptions of the regression model. If $cov(p_3, w) \neq 0$, the OLS estimator is inconsistent [Greene, 1993]. The solution is to use instrumental variables that are correlated with the p_i, but uncorrelated with w. The IV estimator is consistent and asymptotically normal. We used the rating scale valuations, which were elicited as part of the experiment, as instrumental variables. These were believed to be reasonable instrumental variables, because they are correlated with the standard gamble valuations, because they can be expected to be uncorrelated with the error in the standard gamble answers and because they clearly do not belong in the rank dependent utility equations.

6

Experimental results on the ranking properties of QALYs[1]

Summary

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

6.1 Introduction

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

---

[1] Based on Bleichrodt, H. and M. Johannesson, "Standard gamble, time trade-off and rating scale: Experimental results on the ranking properties of QALYs" (submitted for publication).

QALYs are calculated by adjusting life years for the quality of life in which they are spent. Health states are assigned a quality weight that lies between 0 and 1. Three principal methods exist to estimate these quality weights based on the preferences of individuals:[2] the rating scale (RS), the time trade-off (TTO) and the standard gamble (SG).[3] Empirical studies have produced evidence that these three methods elicit different quality weights, the general pattern being that, given common scaling, the standard gamble elicits higher quality weights than the time trade-off which in turn elicits higher weights than the rating scale.[4] The worrying implication of these findings is that QALY based decision making may lead to different policy recommendations depending on which of the three methods is used to elicit the quality weights.

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

Empirical studies comparing the three assessment methods can be broadly divided into two categories: studies that take one of the methods (typically the standard gamble) as the standard against which the performance of the other methods is judged and studies that only compare the quality weights elicited by the three methods without drawing any inferences about their relative performance.

---

[2] In this chapter we will use the term "quality weights." Other terms are in use as well, for example "preference scores." In the context of this chapter these terms are equivalent.

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

[4] Cf. e.g. Torrance (1976), Wolfson et al. (1982), Read et al. (1984), Hornberger et al. (1992)

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

6.2 Standard gamble, time trade-off and rating scale

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

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

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

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

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

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

6.3 Design of the experiment

The aim of the experiment is to examine the relative performance of SG-QALYs, TTO-QALYs and RS-QALYs in terms of their ability to predict individual preferences over health profiles. Respondents were 80 students from the Stockholm School of Economics and 92 students from the Erasmus University Rotterdam. The students were paid approximately $15 for their participation in the study. The experiment was carried out in different sessions lasting approximately one hour with on average ten individuals per session. The procedure followed in each session was first to explain the task to respondents, then to ask respondents to perform the specific task and then to explain the next task. A "master version" of the experiment was designed in English. This "master version" was subsequently translated into Swedish and Dutch. Before drafting the final version, we tested the questionnaire extensively both in Stockholm and in Rotterdam using faculty staff members as respondents.

We selected eight health states to be included in the questionnaire. The health states were taken from the Maastricht Utility Measurement Questionnaire, a slightly adapted version of the McMaster Health Utility Index [Bakker et al., 1994; Rutten-van Mölken et al., 1995]. The selected health states correspond to commonly occurring types of back pain and rheumatism. Health states consist of four attributes: general daily acivities, self care, leisure activities and pain. The attributes and the levels of the attributes are shown in table 6.1. The health states were indicated by capital letters and were described on a set of cards, which were handed out to respondents at the beginning of each session. Health state D, which is relevant for the analysis of this chapter is described in table 6.2.

Table 6.1: The multi-attribute health status classification system used in the experiment.

General daily activities

- Able to perform all tasks at home and/or at work without problems
- Able to perform all tasks at home and/or at work, albeit with some difficulties
- Not able to perform some tasks at home and/or at work
- Not able to perform many tasks at home and/or at work
- Not able to perform any task at home and/or at work

Self care

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

Leisure activities

- Able to perform all types of leisure activities without difficulties
- Able to perform all types of leisure activities, albeit with some difficulties
- Not able to perform certain types of leisure activities
- Not able to perform many types of leisure activities
- Not able to perform any type of leisure activities

Pain and/or other complaints

- No pain and/or other complaints.
- Now and then light to moderate pain and/or other complaints
- Often light to moderate pain and/or other complaints
- Often moderate to severe pain and/or other complaints
- Always severe pain and/or other complaints

Table 6.2: Description of health state D used in the experiment.

- Unable to perform some tasks at home and/or at work
- Able to perform all self care activities (eating, washing, dressing) albeit with some difficulties
- Unable to participate in many types of leisure activity
- Often moderate to severe pain and/or other complaints

We divided the questionnaire into different sections. The first substantive task respondents were confronted with was the ranking of six of the health states in terms of desirability to themselves. For reasons not related to the present study the 6 health states to be ranked varied per session. However, for every respondent health state *D* was included in the ranking task. After the ranking task respondents were asked to locate the health states on a rating scale. In the third section respondents answered three time trade-off questions. All respondents answered a question where one of the alternatives was 30 years in health state *D*. The answer to this time trade-off question is used in the subsequent analysis of this chapter as the time trade-off quality weight for health state *D*. Value elicitation was on a line with numbers of years in full health. Respondents were encouraged to indicate first the values of *X*, the number of healthy years, for which they definitely preferred 30 years in health state *D*, then the values of *X* for which they definitely preferred *X* years in full health and finally those values of *X* for which they found it hard to choose between the alternatives. Respondents were explained that they could indicate a range of values for *X* for which they found it hard to choose, but they were encouraged to make this range as small as possible. This format was adopted to allow respondents to express imprecision of preferences [Dubourg et al., 1994]. Trading off life years is a task respondents are relatively unfamiliar with and their preferences may be somewhat imprecise. In our format we attempted to take this imprecision of preference into account. For individuals who indicated a range of values for *X*, we used the mid-point of this interval as their time trade-off quality weight for health state *D*.

Section four consisted of three standard gamble questions. All versions of the questionnaire contained a question where 30 years in health state *D* was the certain option. Respondents' answers to this question were used in the analysis. Probability elicitation was by means of a line of values for the probability of successful treatment (full health). Next to this line a line was drawn with the complementary probability of failure of treatment (immediate death). This display was chosen in an

attempt to control for a potential framing bias: only displaying the probability of successful treatment might cause individuals to focus on successful treatment, not sufficiently taking into account the probability of failure of treatment. Psychological evidence on the influence of reference effects on choice is abundant [e.g. Tversky and Kahneman, 1991]. Similar to the time trade-off question, an attempt was made to take imprecision of preferences into account. Respondents were asked first to indicate those values of p, the probability of successful treatment, for which they definitely preferred the certain option, then those values of p for which they definitely preferred the treatment option (gamble) and finally those values of p for which they found it hard to choose. For individuals who indicated a range of values for p, we used the mid-point of this interval as their standard gamble quality weight.

In section five the respondents were asked to rank seven health profiles, i.e. combinations of quality of life and quantity of life, in terms of desirability to themselves. The ranking exercise was intendeded to measure individual preferences for health states directly. The health profiles were printed on a set of cards which were distributed together with the questionnaire and the set of health states. This ranking task was similar to the task in section 1, except that the objects to be ranked were health profiles rather than health states. The seven health profiles are described in table 6.3. The seven health profiles differed over 20 years. After 20 years all profiles resulted in death. Profiles lasting less than 20 years resulted in earlier death. In the case of mixed health profiles consisting both of years in full health and years in D, the years in full health always came first. It was expected, and confirmed in the pilot sessions, that profiles of decreasing quality of life are more intuitive to respondents than profiles of increasing quality of life.

Table 6.3: The seven health profiles included in the experiment.

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

To be able to compare the relative performance of SG-QALYs, TTO-QALYs, and RS-QALYs in terms of the direct ranking of the health profiles, health profiles are

needed that will be ranked differently if the QALY weight differs between the methods. The seven health profiles were therefore selected with the intention in mind to produce different rankings for SG-QALYs, TTO-QALYs and RS-QALYs. Table 6.4 uses hypothetical weights between 0.5 and 1.0 to show how the implied QALY ranking of the profiles differs for different quality weights. The experiment was designed with a time trade-off quality weight of about 0.7 in mind. Because empirical evidence has indicated that the standard gamble generally results in higher quality weights than the time trade-off which in turn gives higher weights than the rating scale, we hoped to create the conditions under which the implied rankings for SG-QALYs, TTO-QALYs and RS-QALYs were likely to differ. In the pilot test the average time trade-off weight for health state *D* was about 0.69. The standard gamble weight was 0.76 and the rating scale weight was 0.36. Health state *D* appeared to be a good candidate to test the ranking properties and was therefore selected. Note that the ranking of some of the health profiles in table 6.4 will by definition be the same for all the QALY measures. Profiles 2-5 for instance consist of profiles of varying duration in full health followed by death. We included profiles 2-5 in the experiment to ensure variation in the ranking of all profiles over a rather wide range of quality weights. Without enough variation in the rankings over the range of quality weights we might not detect differences between the methods in correlation with the direct ranking. If the three methods would lead to different quality weights, but the ranking would not vary over this range of quality weights our analysis would not be informative.

Table 6.4: Implied QALY rankings of the seven health profiles for different quality weights (W) with no discounting[a]

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

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

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

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

6.4 Methods of analysis

The seven health profiles were translated for each respondent to SG-QALYs, TTO-QALYs and RS-QALYs on the basis of the elicited quality weights. The predicted rankings of the health profiles were then compared with the direct ranking which was elicited in section 5 of the questionnaire. We examined the predictive power of SG-QALYs, TTO-QALYs and RS-QALYs both at the individual and at the societal level. At the individual level, we compared for each individual the predicted ranking of the health profiles by each of the three models with the direct ranking. To assess the strength of the association we calculated for each individual and for each method the Spearman rank correlation coefficient between the predicted ranking and the direct ranking. These rank correlation coefficients were then averaged over all individuals. The Spearman rank correlation coefficient is a non-parametric

technique which is applicable to ordinal data. Given that the direct ranking data were ordinal, parametric correlation coefficients could not be applied. The Spearman rank correlation coefficient lies between -1 and 1, a higher value indicating stronger positive association between the ranks, a value of zero indicating no association. The QALY measure with the highest average Spearman rank correlation coefficient is most closely associated with the direct ranking of the health profiles.

For the analysis at the societal level we had to aggregate individual preferences into social preference. A problem arises here because of the nature of our data. The direct ranking only provides information with respect to the *ordering* of profiles. No cardinal information is available. Arrow (1951,1963) has proved that it is impossible to construct a social ordering from individual orderings, that satisfies four "very mild looking conditions" [Sen, 1970][6]. We examined two simple social choice rules, each violating one of Arrow's conditions. First, we applied the method of majority voting. Ranking one profile above another was interpreted to be a vote in favour of the former. We constructed a social preference relation from an examination of the votes between every possible pair of profiles. A problem with the social preference relation thus constructed is that it need not be transitive. Therefore we also constructed a social preference relation based on the Borda rule. The Borda rule assigns points to profiles corresponding to the ranks of the profiles and then sums these points over all individuals. The points assigned to a profile were set equal to the rank of the profile. That is, we assigned points in descending order, i.e. a lower number meaning "more preferred." The Borda rule satisfies transitivity, but violates the condition Arrow refers to as "independence of irrelevant alternatives." According to this condition social preference between two alternatives should not be affected by a third alternative. The method of majority voting and the Borda rule each satisfy a different subset of Arrow's conditions. The union of these sets consists of Arrow's conditions.

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

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

themselves (see Labelle and Hurley (1992) for a discussion of the potential importance of preferences over outcomes for others).

6.5 Results

6.5.1 Disparity between the methods

Differences in ranking performance can only occur if the three methods produce different weights. Table 6.5 shows that the mean quality weights for standard gamble, time trade-off and rating scale indeed differ significantly. Differences are significant at the 0.1% level, both for the Dutch, the Swedish and the total sample. The difference between the Dutch and the Swedish sample is not significant ($p > 0.10$) for the time trade-off and the standard gamble quality weights. However, for the rating scale the difference is significant at the 10 % level ($p \approx 0.09$). Compared to the results of the pilot study the standard gamble weights and the time trade-off weights are approximately 0.10 lower, whereas the rating scale weight is slightly higher.

Table 6.5: Mean SG, TTO and RS quality weights for health state D. Standard errors within parentheses.

Method	Dutch sample	Swedish sample	Total sample	Difference NL-S
RS	0.3867[b]	0.4274[b]	0.4056[b]	p = 0.092
	(0.0149)	(0.0191)	(0.0199)	
TTO	0.5958[a,b]	0.5575[a,b]	0.5780[a,b]	n.s.
	(0.0238)	(0.0237)	(0.01684)	
SG	0.6786[a]	0.6620[a]	0.6709[a]	n.s.
	(0.0279)	(0.0257)	(0.0191)	

a: significantly different from RS at 99% confidence level.

b: significantly different from SG at 99% confidence level.

6.5.2 Spearman rank correlation coefficients

Table 6.6: : Mean Spearman rank correlation coefficients between the direct ranking and the predicted rankings of RS-QALYs, TTO-QALYs, and SG-QALYs. Standard errors within parentheses. No discounting.

Comparison	Dutch sample	Swedish sample	Total sample	Difference NL-S
RS - Direct	0.7208	0.7932	0.7545	p = 0.088
	(0.033)	(0.027)	(0.022)	
TTO - Direct	0.8194[b,c]	0.8669[b,c]	0.8415[a,c]	n.s.
	(0.027)	(0.018)	(0.017)	
SG - Direct	0.6891	0.7684	0.7259	p = 0.062
	(0.036)	(0.027)	(0.023)	

a: significantly different from RS at 99% confidence level.

b: significantly different from RS at 95% confidence level.

c: significantly different from SG at 99% confidence level.

Table 6.6 displays the mean of the Spearman rank correlation coefficients between the direct rankings and the rankings predicted by RS-QALYs, TTO-QALYs and SG-QALYs respectively. Both in the Dutch and in the Swedish sample TTO-QALYs are significantly stronger correlated with the direct ranking than RS-QALYs and SG-QALYs. The difference between RS-QALYs and SG-QALYs is not significant even though the mean rank correlation coefficient is higher for RS-QALYs in both samples.

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

Table 6.6 was constructed under the assumption of no discounting. Table 6.7 shows the results for a discount rate of 5%. This is the situation most frequently encountered in cost utility analyses.

Table 6.7: Mean Spearman rank correlation coefficients between the direct ranking and the predicted rankings of RS-QALYs, TTO-QALYs, and SG-QALYs. Standard errors within parentheses. 5 % discounting.

Comparison	Dutch sample	Swedish sample	Total sample	Difference NL-S
RS - Direct	0.7058	0.7795	0.7401	n.s.
	(0.035)	(0.030)	(0.023)	
TTO - Direct	0.7886[b,c]	0.8481[b]	0.8162[a,c]	p = 0.086
	(0.028)	(0.020)	(0.018)	
SG - Direct	0.7297	0.8274	0.7752	p = 0.022
	(0.033)	(0.026)	(0.023)	

a: significantly different from RS at 95% confidence level.

b: significantly different from RS at 90% confidence level.

c: significantly different from SG at 90% confidence level.

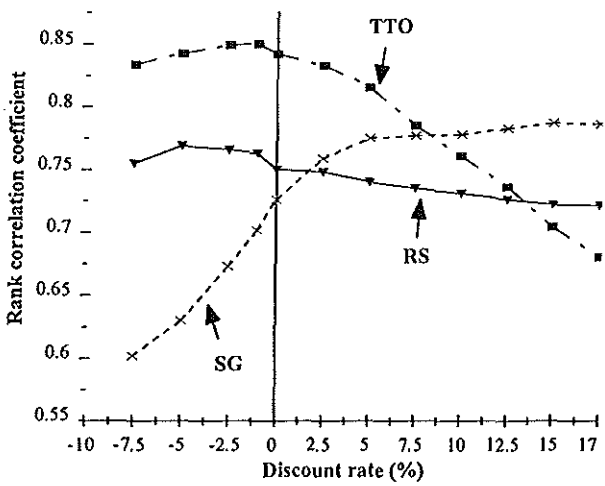
A comparison between tables 6.6 and 6.7 reveals that 5 % discounting reduces the rank correlation coefficients for rating scale and time trade-off and increases the rank correlation coefficient for the standard gamble. Even though the difference is less significant, the rank correlation with the direct ranking is still higher for TTO-QALYs than for RS-QALYs and SG-QALYs. For the latter two, the difference between the mean rank correlation coefficients is not significant.

A comparison between tables 6.6 and 6.7 further shows that for the standard gamble the mean rank correlation coefficient increases with 5% discounting. The difference between the mean rank correlation coefficients is highly significant *($p < 0.001$)*. Figure 1 shows for the total sample that for discount rates higher than 5%, the rank correlation coefficient for SG-QALYs increases even more. For the total sample the maximizing discount rate for SG-QALYs is 15.5% *(rcc = 0.789)*. For the Swedish sample the maximum rank correlation coefficient for SG-QALYs is attained at a discount rate of approximately 9% *(rcc = 0.835)*. For the Dutch sample the maximum rank correlation coefficient for SG-QALYs is attained at a discount rate of approximately 16 % *(rcc = 0.760)*.

For the rating scale and the time trade-off we observe the opposite pattern: 5% discounting has a decreasing impact on the mean rank correlation coefficient. For the rating scale the difference is significant at the 5% level. For the time trade-off the difference is significant at the 0.1% level. Figure 1 shows that the maximizing discount rate for RS-QALYs is approximately -5 % *(rcc = 0.769)* for the total

sample. For the Swedish and the Dutch samples the maximizing discount rates for RS-QALYs are -1% *(rcc = 0.804)* and -5% *(rcc = 0.756)* respectively. For TTO-QALYs the maximizing discount rate is approximately -1% *(rcc = 0.850)*. For the Dutch and Swedish samples the maximizing discount rates are -5% *(rcc = 0.829)* and -1% *(rcc = 0.876)* respectively.

Figure 6.1: The relationships between discount rates and rank correlation coefficients.



6.5.3 Majority voting

Table 6.8 summarizes the evidence on majority voting. Profiles are shown in decreasing order of preference according to majority voting. The first column shows the results for the direct ranking exercise. Columns 2-4 show the results for the three methods with no discounting and columns 5-7 show the results for the three methods with 5% discounting. The number in parentheses shows the percentage of respondents who favour the profile over the profile coming next in preference. For example, the column "direct" shows that in the direct ranking task profile 2 came out as most preferred, profile 3 as second most preferred, profile 4 as third most preferred et cetera. Profile 2 was preferred by all respondents to profile 3, profile 3

was preferred by all respondents to profile 4, and profile 4 was preferred by 71% of respondents to profile 6.

Table 6.8 shows that for no discounting, the ranking predicted by TTO-QALYs corresponds most closely to the ranking being given: the predicted rank order of the profiles is similar to the rank order elicited in the direct ranking task. Moreover, the predicted percentage of respondents "voting" in favour of a profile is in most cases quite similar to the percentage of respondents "voting" in favour of a profile according to direct ranking. To obtain insight in the correspondance of the proportions voting in favour of a particular profile, we calculated the correlation coefficients between the proportion of votes in favour of profile *i* over profile *j* based on the direct ranking and the predicted proportion of votes in favour of profile *i* over profile *j* based on each of the three methods. These correlation coefficients are shown in the bottom rows of table 6.8.[7] In the situation of no discounting the correlation coefficient is significantly higher for the time trade-off than for the rating scale and the standard gamble at the 0.1% significance level.[8] The difference between rating scale and standard gamble is not significant. In the situation of 5% discounting the correlation coefficient is highest for the standard gamble. Only the difference between standard gamble and rating scale is significant ($p = 0.022$).

Table 6.8 also suggests an explanation why SG-QALYs and RS-QALYs are less consistent with the results of the direct ranking exercise than TTO-QALYs. The explanation we suggest is that the standard gamble assigns too much weight to health state *D* compared to the weight implied by direct ranking. The rating scale does not assign enough weight to health state *D* compared to the weight implied by direct ranking. To illustrate the first claim made, that the standard gamble assigns too much weight compared to direct ranking, compare for example profiles 1 and 5. The difference between profiles 1 and 5 is that profile 1 offers more life years than profile 5, but these life years are spent in a lower quality of life. Profile 1 will be preferred to profile 5 if the utility gain of 8 additional years in health state D more than compensates the utility loss of spending the first 12 years in health state D rather than in full health. For example, in the situation of no discounting, profile 1 will be preferred to profile 5 if $8*[U(D)-U(immediate\ death)] > 12*[U(full$

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

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

health)-U(D)].[9] Or, given the scaling of U, if $U(D) > 0.6$. Table 6.8 shows that according to the standard gamble profile 1 is preferred to profile 5 by a majority of respondents In fact 70.9% preferred profile 1 to profile 5 according to SG-QALYs. Thus 70.9% of the respondents assigned a quality weight greater than 0.6 to health state D by the standard gamble. However, in the direct ranking a majority of respondents (53%) ranked profile 5 above profile 1, which implies that only 47% of the respondents assigned a quality weight greater than 0.6 to health state D in the direct ranking task. Thus, compared to the direct ranking the standard gamble gives too much weight to health state D.

Table 6.8: Majority vote ordering of profiles for the direct ranking of profiles and the ranking predicted by RS-QALYs, TTO-QALYs, and SG-QALYs with 0% and 5% discounting. The proportions of individuals preferring a profile to the next profile in the ordering are shown within parentheses.[1]

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

[1] For SG-QALYs with no discounting profiles 4 and 7 were equivalent according to majority voting (i.e. the proportion of respondents "voting for" 4 over 7 was exactly 50%). The table shows that 54% were predicted to prefer profile 1 to profile 4 and 59% were predicted to prefer profile 1 to profile 7. The 55% vs 59% given in parenthesis similarly means that 55% were predicted to prefer profile 4 to profile 6 and 59% were predicted to prefer profile 7 to profile 6.

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

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

The hypothesis that the rating scale assigns too low a weight to health state D compared to the weight implied by the direct ranking can be seen for example by comparing profiles 5 and 6. With no discounting a similar calculation exercise as above shows that profile 6 will be preferred to profile 5 if $U(D) > 0.5$. The direct ranking exercise revealed that a majority of respondents preferred profile 6 to profile 5 (64.6%). That is, 64.8% of the respondents assigned a weight greater than 0.5 to health state *D* in the direct ranking task. However, in the rating scale task only 25% of the respondents assigned a weight greater than 0.5 to health state *D*. Obviously, for the rating scale positive discounting only makes things worse. With 5% discounting profile 6 will be preferred to profile 5 if $U(D) > 0.55$. Thus respondents who give a quality weight between 0.50 and 0.55 to health state *D* are now predicted by RS-QALYs to prefer profile 5.

Table 6.9: Correlation coefficients for the two samples for the majority vote ordering of profiles.

Sample (% disc.)	RS	TTO	SG
NL (0)	0.879	0.979[a,d]	0.806
S (0)	0.924	0.982[b,d]	0.862
NL (5)	0.856	0.932	0.928
S (5)	0.894	0.948	0.981[a]

a: significantly different from RS at 99% confidence level.
b: significantly different from RS at 95% confidence level.
c: significantly different from RS at 90% confidence level.
d: significantly different from SG at 99% confidence level.

Table 6.9 displays the results for the Swedish and Dutch samples separately. We observe a similar pattern as for the total sample. Correlation coefficients are

without exception higher in the Swedish sample. However, the difference between the two samples is only significant for SG (5%) *($p = 0.000$)*.

6.5.4 Borda rule

Table 6.10 shows social preferences according to the Borda rule for the total sample. Recall from section 4 that in a comparison of profiles, the profile with the lowest score is preferred. Thus for the direct ranking exercise the preference order implied by the Borda rule is $2>3>4>6>7>1>5$.

Table 6.10 shows that in the situation where no discounting is applied the ranking predicted by TTO-QALYs is most closely related to the direct ranking. Table 6.10 confirms the pattern we already observed with respect to table 6.8. The standard gamble assigns too high a quality weight to health state D compared to the weight implied by the direct ranking exercise. Positive discounting mitigates this effect: with 5% discounting SG-QALYs more closely reflect the direct ranking. The rating scale assigns too low a weight to health state D compared to the weight implied by the direct ranking exercise. Positive discounting reinforces this effect.

Table 6.10: Mean Borda scores for the health profiles. The first column shows the borda scores based on the direct ranking of profiles and the following columns shows the borda scores based on the predicted ranking of profiles for RS-QALYs, TTO-QALYs, and SG-QALYs with no discounting and 5% discounting.

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

Applications of utility theory in the economic evaluation of health care

The final three rows of table 6.10 report for the total sample, the Dutch sample and the Swedish sample the correlation coefficients of the Borda scores assigned by the three methods with the Borda scores calculated on the basis of the direct ranking.[10] These confirm the remarks made above: the scores predicted by TTO-QALYs are most closely related to the scores resulting from the direct ranking in the case of no discounting. In the case of 5% discounting, the scores predicted by SG-QALYs are most closely related to the scores predicted by direct ranking.

6.6 Concluding remarks

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

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

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

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

It could be argued that even though the correlation between predicted and direct ranking was statistically significantly higher for TTO-QALYs than for RS-QALYs and SG-QALYs, the differences between the methods are not economically important (for a discussion of the distinction between statistical significance and economic importance see McCloskey, 1983). This argument is based on the observation that all methods performed well, because the correlation coefficients were rather high for all three methods by the standards usually used to judge the size of correlation coefficients [cf. Landis and Koch, 1977]. The fact that all correlation coefficients are rather high would also imply support for the use of additively separable utility models in health.

However, one can object against the above argument. It is difficult to interpret the absolute size of the correlation coefficients in the context of this study and to judge the size of the correlation coefficients by the usual standards. The reason is that the predicted ranking of some of the profiles will by definition be the same as the direct ranking as long as individual preferences satisfy monotonicity with respect to years in full health (i.e. preferences between profiles 2,3,4 and 5 are obvious) and as long as the individual prefers years in full health to years in health state D (i.e. profile 2 will always be preferred to profiles 6 and 7 and profile 3 will always be preferred to profile 6). To illustrate the impact on the size of the correlation coefficients of profiles for which the ranking is obvious, we redid the analysis using only the two combinations of health profiles for which the ordering was not obvious beforehand. These combinations are profiles 1,4,6 and 7 and profiles 1,5,6 and 7. For the analysis including only profiles 1,4,6 and 7 the mean Spearman rank correlation coefficient between predicted and direct ranking is 0.55 for TTO-QALYs, 0.46 for RS-QALYs and 0.28 for SG-QALYs.[11] Translated to the classification scheme of Landis and Koch (1977) these rank correlation coefficients would classify as "moderate" for TTO-QALYs and for RS-QALYs and as "fair" for SG-QALYs. For the analysis including only the profiles 1,5,6 and 7 the mean Spearman rank correlation coefficient between predicted and direct ranking is 0.54 for TTO-QALYs, 0.30 for RS-QALYs and 0.35 for SG-QALYs.[12] According to the

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

[12] The difference between the mean rank correlation coefficients for TTO-QALYs and for RS-QALYs and between the mean rank correlation coefficients for TTO-QALYs and for SG-QALYs

classification scheme of Landis and Koch (1977) the correlation is "moderate" for TTO-QALYs and "fair" for RS-QALYs and for SG-QALYs. The pattern is for both combinations quite similar to the pattern we observed when all health profiles were included in the analysis except that the mean rank correlation coefficients are lower. The differences between the mean rank correlation coefficients are larger and the correlation coefficients do not all fall in the same class according to the classification scheme of Landis and Koch (1977). This suggests that the differences between the methods are not only statistically significant, but also meaningful.

As far as we know this study is the first which compares the performance of RS-QALYs, TTO-QALYs and SG-QALYs in terms of direct ranking. Apart from offering some tentative conclusions, this chapter also raises various questions which may be addressed in future research. First, we measured individual preferences by direct ranking of the 7 profiles simultaneously. A different procedure would be to confront individuals with all possible pairs of profiles and to construct a preference ordering from these answers. It is not clear a priori whether the two approaches give identical results. For example, our approach excluded intransitivities. The pairwise approach on the other hand might lead to intransitivities. Second, the approach we used in the time trade-off and standard gamble questions is close to the ping-pong approach favoured by many researchers in the field, but it is not exactly similar. Moreover, to accommodate imprecision of preferences we allowed respondents to indicate ranges of values. Although we do not believe that our slightly different procedures have affected the results, it may be worth investigating the sensitivity of the results to this difference in approach. Third, we used only one health state and only a limited number of profiles for which the ranking was not obvious. It may be that the time trade-off is a useful heuristic for a number of health states, but that it does not work equally well for all health states [cf. Stalmeier et al., 1995]. It is worth redoing the analysis using different health states and profiles. Fourth, the fact that we used group sessions rather than individual sessions may have decreased the care with which some individuals answered the questionnaire. This may in particular have affected the standard gamble responses. The standard gamble is generally considered to be the most complicated method of the three. On the other hand, as can be seen from table 6.5, the pattern of differences in quality weights between the methods is similar to that observed in other studies. Fifth, it is possible that the performance of the methods is affected by the ordering of the tasks in the experiment. All respondents were first asked to perform the rating scale task,

---

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

then the time trade-off task and finally the standard gamble task. The reason we opted for this order was that in general the rating scale is considered the easiest method to answer and the standard gamble the most complicated. However, it may be that during the experiment respondents became more aware of their "true" preferences and thus the higher consistency of TTO-QALYs over RS-QALYs may simply be a consequence of the order in which the tasks were performed. Even though at the end of the experiment we urged respondents to carefully read through their responses again and to make changes where they thought appropriate, ordering effects may have affected the results. Future experimental studies may wish to randomize the order of the tasks or, alternatively, respondents may be asked to perform only one task.

Finally, two notes of warning are worth making. First, we interpreted QALYs as a utility model. Even though this appears to be the most common interpretation of QALYs (for example the recent debate on the merits of QALYs versus healthy-years equivalents [Buckingham, 1993; Culyer and Wagstaff, 1993; Gafni, Birch and Mehrez, 1993; Loomes, 1995; Johannesson, 1995; Bleichrodt, 1995] focused on the question of the consistency of QALYs with individual preferences), there are other interpretations, as we remarked before. Second, our results only bear relevance for the descriptive validity of the various QALY models. It may be that for normative/prescriptive reasons, which are more relevant in health economics and medical decision making, one wishes to stick to SG-QALYs. Moreover, if QALYs are intended as decision aids to prescribe individual choices, the paradoxical result emerges that once a model corresponds perfectly with direct choices the model looses its significance for prescriptive purposes. In the case of perfect correspondence one could simply let individuals choose intuitively and no decision-aiding analysis would contribute anymore. Predictive models can only be of use if they deviate somewhat from actual choice. The question then obviously is how much we allow our measures to deviate from actual choice. This is a question that may be picked up in future research.

7

Time Preference, The Discounted Utility Model and Health[1]

Summary

The constant rate discounted utility model is commonly used to represent intertemporal preferences in health care program evaluations. This chapter examines the appropriateness of this model, and argues that the model fails both normatively and descriptively as a representation of individual intertemporal preferences for health outcomes. Variable rate discounted utility models are more flexibile, but still require restrictive assumptions and may lead to dynamically inconsistent behaviour. The chapter concludes by considering two ways of incorporating individual intertemporal preferences in health care program evaluations that allow for complementarity of health outcomes in different time periods.

7.1 Introduction

Constant rate discounted utility models are commonly used to represent intertemporal preferences in health care program evaluation. The debate mainly centers around the question of what rate of discount to use. Little attention has been paid to the appropriateness of the constant rate discounted utility model as such. The axioms underlying the individual preference structure to fit impatience [Koopmans, 1960], time perspective [Koopmans et al, 1964] and the discounted utility model both for a single outcome [Fishburn and Rubinstein, 1982] and for (infinite) sequences of outcomes [Koopmans, 1972] can be found in the economic literature. The general impression from this literature is that the discounted utility model[2] is far from realistic. This impression has been confirmed by empirical studies concerning time

[1] Based on Bleichrodt, H. and A. Gafni, 1995, "Time preference, the discounted utility model and health," Journal of Health Economics (in press).

[2] From now on discounted utility model will stand for constant rate discounted utility model unless otherwise stated.

preference. These studies display a number of anomalies, that are robust and do not require ingenious experimental designs to be revealed.[3]

This chapter examines the appropriateness of the discounted utility model as a description of an individual's intertemporal preferences for health outcomes. The analysis has immediate relevance for the appropriateness of the use of the discounted utility model in the context of eocnomic evaluations where the social discount rate is assumed to be based on the aggregate of individuals' intertemporal preferences [e.g., Redelmeier and Heller, 1993; Weinstein, 1993]. The conditions that the model imposes on the individual preference structure are derived and their restrictiveness is assessed. Both the case where the preference relation is defined over health outcomes and the case where the preference relation is defined over lotteries over health outcomes are addressed. We argue that in neither situation does the discounted utility model provide a good description of an individual's intertemporal preferences for health outcomes.The argument that the discounted utility model may not hold descriptively, but should be adopted because of its normative appeal will be considered but ultimately rejected.

It has been argued that the rejection of a constant rate of discount calls for the use of a model with a discount rate that is variable [see for example Olsen, 1993b]. By examining the axiomatic structure of the model and by means of an example, we show that using a variable rate discounted utility model does not solve all problems of the constant rate discounted utility model and creates a problem of its own: it may entice the individual to behave in a dynamically inconsistent way.

In this chapter we are concerned mainly with individual intertemporal preferences for sequences of health outcomes. One might argue that an individual's intertemporal preferences are of no interest in health care program evaluations given that health care program evaluation should be based on an appropriately selected social rate of discount. But when the social discount rate is to be based on the aggregate of the individual intertemporal preferences, as has often been advocated in the case of program evaluation, this argument runs into problems. It is not clear why an aggregate concept should satisfy a model that is violated by its constituent parts. On the other hand defining the social rate of discount without taking into account the individual's intertemporal preferences raises the question - what should the foundation of the social rate of discount be? It has been argued that one should select the appropriate market rate of interest corrected for tax distortions. However,

---

[3]See for example Loewenstein (1987), Loewenstein (1988), Loewenstein and Prelec (1991), Loewenstein and Prelec (1992), Loewenstein and Prelec (1993), Loewenstein and Thaler (1989), and Thaler (1981a). For examples of violations of the discounted utility model with health outcomes see Olsen (1993a), and Redelmeier and Heller (1993).

correcting for tax distortions is far from straightforward and the relationship between the market rate of interest and the social rate of time preference is further distorted by the internationalization of capital markets [Lind, 1990]. Also, ignoring individual intertemporal preferences might be undesirable for reasons of consistency. Considerable attention is being given to the development of methods to elicit individuals' preferences for health outcomes. Because health outcomes have a time dimension inextricably bound to them, we cannot ignore individual intertemporal preferences in valuing them.

The structure of the chapter is as follows. Section 2 derives the discounted utility model when the preference relation is defined over health outcomes under certainty. Section 3 derives the discounted utility model when the preference relation is defined over lotteries on health outcomes (i.e., under risk). Both sections are technical. In section 4 the axioms underlying the discounted utility model are discussed from a normative point of view. Section 5 presents descriptive evidence concerning factors affecting individual intertemporal preferences. In section 6 we discuss the argument that a variable rate discounted utility model should be used to model individual intertemporal preferences for health. Section 7 contains concluding remarks and considers two alternative approaches to incorporate individual intertemporal preferences in the evaluation of health care programs. Proofs of the various results presented in the chapter appear in the appendices.

7.2 Intertemporal preferences under certainty

7.2.1 Preliminaries

This subsection introduces notation and structural assumptions. For more details the reader is referred to the appendices. The chapter deals with an individual decision maker who has a preference relation $\succsim$, meaning "at least as preferred as", over a set X of health profiles. A typical element of X is $(x_1, x_2, \ldots, x_T)$ where x_i denotes health status in period i and T denotes the remaining number of years the individual decision maker will live until death. The x_i are elements of identical one-period sets of health outcomes A.

We assume the preference relation $\succsim$ over X to be a continuous weak order. A weak order is (i) complete: the individual decision maker can rank all health profiles, and (ii) transitive: if the individual decision maker considers profile x to be at least as good as profile y $(x \succsim y)$ and profile y to be at least as good as profile z $(y \succsim z)$, then the individual should also consider profile x to be at least as good as profile z $(x \succsim z)$. Strict preference and indifference are denoted by $\succ$ and $\sim$ respectively.

Elements of X, the set of health profiles, are denoted by Roman characters x,y, etc. Elements of A, the one period sets of health outcomes, are denoted by Greek characters α,β etc. Constant alternatives are alternatives that give health outcome α in every period, and are denoted by α_χ. We write $x_{.i}\alpha$ to denote the health profile x with x_i replaced by health outcome α. Similarly, $x_{.i,j}\alpha,\beta$ denotes health profile x with x_i replaced by α, and x_j replaced by β.

7.2.2 Preference conditions

Definition 7.1: The preference relation $\succsim$ is called coordinate independent (CI) if
$(x_{.i}\alpha) \succsim (y_{.i}\alpha) \Leftrightarrow (x_{.i}\beta) \succsim (y_{.i}\beta)$ *for all* x,y,i,α,β

The idea underlying CI is that if two alternatives have an identical health outcome in a certain period (have a coordinate in common), then the preference between these alternatives should be unaffected when that common health outcome is changed into another common health outcome. CI is also known by other names in the literature: e.g., independence (Debreu, 1960; Krantz et al., 1971), mutual preferential independence (Keeney and Raiffa, 1976).

Definition 7.2: The preference relation $\succsim$ is called cardinally coordinate independent (CCI) if for all x,y,v,w,a,b,g,d,j *and* i,
$(x_{.i}\alpha) \precsim (y_{.i}\beta)$ & $(x_{.i}\gamma) \succsim (y_{.i}\delta)$ & $(v_{.j}\alpha) \succsim (w_{.j}\beta)$ *imply* $(v_{.j}\gamma) \succsim (w_{.j}\delta)$.

The intuition behind this condition is as follows. Suppose α is preferred to β and γ is preferred to δ. One might say that in period i, the strength of preference of α over β is smaller than the strength of preference of γ over δ, since trading off β for α is not sufficient to compensate for getting x rather than y in all other time periods, whereas trading off δ for γ is sufficient. By CCI, if in period j the strength of preference of α over β is sufficient to compensate for getting v rather than w in all other time periods, then trading-off δ for γ is also sufficient. CCI establishes that trade-offs between health outcomes are not contradictory in different periods.

Definition 7.3: The preference order $\succsim$ is called impatient if
$\alpha_c \succsim \beta_c \Leftrightarrow (x_{.i,i+1}\alpha,\beta \succsim x_{.i,i+1}\beta,\alpha)$ *for all* x,α,β

According to definition 7.3, an individual is impatient if he prefers favourable outcomes to occur sooner rather than later. Impatience excludes the possibility that

individuals prefer to postpone favourable outcomes because of the derivation of utility from the anticipation of future favourable outcomes.

Definition 7.4: The preference order $\succsim$ *is called stationary if, for a constant alternative x, there exist health outcomes* α *and* β *such that for all time periods i:*

$(x_{.i}\beta) \sim (x_{.i+1}\alpha)$.

Stationarity has the effect of making the trade-off between health outcome β in time period i and health outcome α in time period $i+1$ invariant with respect to what time period i is. The trade-off between health outcomes occurring at different points in time depends only on the difference in time of occurrence between the health outcomes and not on the exact point in time at which they occur.

We are now ready to state a first theorem.

Theorem 7.1: The following two statements are equivalent:

(i) There exists a unique $0 < \pi \leq 1$, *and a continuous function* $V: A \rightarrow \mathbb{R}$, *increasing up to positive affine transformations, such that the individual preference relation* $\succsim$ *over the set of health profiles X can be represented by*

$$W(x) = \sum_{i=1}^{T} \pi^{i-1} V(x_i) \qquad (1)$$

(ii) The preference relation $\succsim$ *is a continuous weak order, it satisfies CCI, impatience and stationarity.*

The proof of this theorem can be found in the appendix.

7.3 Intertemporal preferences under uncertainty

A widely held view in health economics is that, since risk is an essential element of health decision making, and no appropriate mechanisms exist for spreading the risk, individual attitudes towards risk should be incorporated in the decision making process both at the individual and group level (e.g., Ben Zion and Gafni, 1983). A way to achieve this, following von Neumann and Morgenstern (1953), is to define preferences over lotteries over health outcomes rather than over the health outcomes themselves. We refer to lotteries over health outcomes as risky health outcomes.

7.3.1 Preliminaries

In the context of decision making under risk, the individual preference relation $\succsim_Z$ is defined over the set Z of simple probability distributions (lotteries) over X. Elements of Z are denoted by capital Roman characters, P, Q, etc. Lotteries over A, the set of one period health outcomes, are denoted by $P_{i\_}$, $Q_{i\_}$, etc. P_i and Q_i are marginal probability distribtions. We assume that the preference relation $\succsim_Z$ satisfies the von Neumann Morgenstern (vNM) axioms. These axioms are necessary and sufficient for the existence of a cardinal, real valued function $U:X \rightarrow \mathbb{R}$, the expectation of which represents $\succsim_Z$. It is important to realize that in vNM utility theory Z contains all degenerate probability distributions assigning probability one to an alternative. This induces a preference relation $\succsim$ over X. Note that U represents $\succsim$.

7.3.2. Preference conditions

In deriving the discounted utility representation, we make maximal use of the preference conditions defined in section 2. An alternative approach would be to reformulate these conditions over risky health outcomes rather than over health outcomes [e.g. Fishburn, 1970 (section 11.4)]. In our opinion defining preference conditions over the set of risky health outcomes makes the conditions less intuitive. We therefore restrict the use of conditions on the set of risky health outcomes to a minimum. However, one assumption on the set of risky health outcomes is necessary in order to relate risky health outcomes and health outcomes.

Definition 7.5: The preference relation $\succsim_Z$ on Z is called <u>additive independent</u> if

$$[P, Q \in Z,\ P_i = Q_i \text{ for } i = 1,..,T] \Rightarrow P \sim_Z Q$$

Additive independence asserts that preferences over risky health outcomes depend only on the marginal probability of occurrence of each health outcome and not on their joint probability distribution. If two probability distributions result, at each point in time, in the same probability distribution over health outcomes, then by additive independence they should be indifferent.

Now a second theorem can be given.

Theorem 7.2: The following two statements are equivalent:

(i) There exists a unique $0 < \pi \leq 1$, and a continuous vNM utility function $U: A \rightarrow \mathbb{R}$, increasing up to positive affine transformations, such that the individual preference relation $\succeq$ over health profiles can be represented by

$$U(x) = \sum_{t=1}^{T} \pi^{t-1} U(x_t) \qquad (2)$$

(ii) The preference relation $\succeq_z$ over risky alternatives is a weak order, it satisfies vNM independence and Jensen continuity and additive independence. Restricted to degenerate probability distributions, $\succeq$ satisfies CCI, impatience and stationarity.

The proof of this theorem can be found in the appendix.

7.4 A normative assessment of the preference conditions

Having identified the preference conditions underlying the discounted utility model, the question emerges of how appealing are these conditions. This section considers whether individuals should behave according to these preference conditions.

Coordinate independence is a strong assumption. It excludes complementarity of health states over times. Therefore phenomena such as coping and maximal endurable time (Sutherland et al., 1984), that depend on sequences of health states cannot be accounted for within the framework of the model. An example may clarify how CI excludes complementarity.

Suppose there are three points in time (three coordinates): $i = 1,2,3$ and three health states: good health (G), mediocre health (M) and poor health (P). Consider two choices: $A = (M_1, G_2, G_3)$ versus $B = (G_1, M_2, G_3)$ and $A' = (M_1, G_2, P_3)$ versus $B' = (G_1, M_2, P_3)$, where M_i stands for mediocre health in time period i. It is conceivable that an individual prefers A to B, because he would rather "get over" mediocre health quickly or because he is averse to changes in his health status. It is also conceivable that the same individual prefers B' to A', because he feels it is easier to cope with P_3 when his health decreases gradually over time. A preference for A over B and for B' over A' is caused by complementarity of health outcomes over time. Both variation aversion and coping relate to sequence effects. CI excludes the combination of A preferred to B and B' preferred to A'. The two choice situations differ only in the common third coordinate and, since by CI common coordinates cannot influence preference it follows that these two choice situations are equivalent.

When the coordinates i are states of the world rather than time points, CI is equivalent to Savage's (1954) sure thing principle that preferences between alternatives should not be influenced by states of nature in which the two alternatives have common outcomes, regardless of what those common outcomes are. This is exactly what CI implies: common coordinates do not influence the preference relation.

The sure thing principle is theoretically less appealing when coordinates are points in time rather than states of nature. The traditional defence of the sure thing principle in the context of decision making under uncertainty [e.g. Samuelson, 1952], that something that never happens should not influence the value of something that actually does take place, does not carry over. In the points of time interpretation all time periods do occur.

It is a common belief in economics that individuals do indeed prefer benefits sooner rather than later, which supports impatience. Also, Olson and Bailey (1981) provide several normative arguments in defence of impatience. However, impatience excludes such effects as anticipation and dread. In the context of health decision making it does not seem irrational to prefer unpleasant events to happen sooner rather than later.

Stationarity lacks normative appeal as the time preference literature acknowledges. For example Fishburn and Rubinstein (1982) claim: "...we know of no persuasive argument for stationarity as a psychologically viable assumption" [p.681]. Similar views have been expressed by Koopmans (1960, 1972). Stationarity requires the passage of time to have no influence on preferences. However, if an individual is indifferent between health improvement A now and health improvement B with a certain time delay x, why should this individual be indifferent between health improvement A in a year's time and health improvement B at time $x + 1$ year?

Finally, additive independence is a strong condition. Additive independence excludes any complementarity of health outcomes in different time periods. For example, it requires that an individual is indifferent between two treatment scenarios A and B, where A results with probability 0.5 in "living 40 years in good health" and with probability 0.5 in "living 40 years in a poor health state, P" and B results with probability 0.5 in "first living 20 years in good health followed by 20 years in P" and with probability 0.5 in "first living 20 years in P followed by 20 years in good health". In both treatment scenarios, in every year the individual has a probability of 0.5 of being in good health and a probability of 0.5 of living in health state P. Therefore, by additive independence, indifference should hold. However, some people could for example prefer treatment A because this gives the prospect of living the rest of their lives in good health, while others might prefer treatment B because this guarantees living 20 years in good health. For a more elaborate discussion of the appropriateness of additive independence in health decision making see Maas and Wakker (1994).

In summary, it appears that no persuasive arguments exist as to why an individual should behave according to the discounted utility model. It has been suggested by several authors [e.g. Weinstein (1993)] that the discounted utility model can be placed normatively on the same footing as the expected utility model. However, the translation of the expected utility model to the time context reduces the appeal of the underlying axioms and the discounted utility model also requires additional, restrictive, axioms.

7.5 A descriptive assessment of the preference conditions

This section considers the descriptive validity of the discounted utility model. First, an overview is given of the various factors that have been identified in empirical work as influencing individual intertemporal preferences. Second, direct empirical evidence is presented on the appropriateness of the discounted utility model in health decision making.

7.5.1 A decomposition of intertemporal preference

Olson and Bailey (1981), following Böhm-Bawerk, identify two "influences" which cause an individual to have a positive rate of time preference: decreasing marginal utility and pure time preference. Furthermore, they mention the influence of uncertainty on intertemporal preferences, but do not predict the sign of this effect. Gafni and Torrance (1984) have translated these effects to the case of a chronic health state. They identify the following three influences: *i)* a quantity effect (decreasing marginal utility of health); *ii)* a gambling effect, a consequence of the presence of uncertainty; and *iii)* a pure time preference effect, reflecting the fact that individuals prefer to receive benefits sooner rather than later.

If present, all three of these effects will be properly handled by the discounted utility model. Decreasing marginal utility and the individual's attitude towards uncertainty are reflected by the shape of the utility function, and the pure time preference effect is incorporated in the discount factor. However, the analysis by Gafni and Torrance shows that separating these three different effects, which seems necessary in order to include them in a credible way in the discounted utility model, may prove to be a cumbersome task.

More recent empirical work[4] suggests that other influences besides the three mentioned above affect intertemporal preferences. Individuals generally prefer increasing profiles to decreasing profiles that are a permutation of these increasing profiles, both for wages [Frank and Hutchens, 1993; Loewenstein and Sicherman, 1991] and for other attributes [Loewenstein and Prelec, 1991], contrary to the predictions of the discounted utility model. Explanations of this fact distinguish sequences from single outcomes. Kahneman and Thaler (1991) identify adaptation and loss aversion as important influences on intertemporal preferences for sequences. Adaptation (also called anchoring) refers to the idea that the individual tends to consider the normal to be neutral, neither good nor bad. Adaptation is the foundation for Scitovsky's (1976) distinction between comforts, which become noticeable only when they are withdrawn, and pleasures, which are noticeable being distinct from the normal. Adaptation plays a central role in Loewenstein and Prelec's (1992) model of intertemporal choice. Streams of outcomes are evaluated as deviations from a reference vector rather than as being incorporated in existing plans. Loss aversion refers to the fact that the value function for losses is steeper than for gains. Adaptation in combination with loss aversion causes changes in the levels of well-being rather than absolute levels of well-being to be the real carriers of value for an individual.

Adaptation and loss-aversion are relevant for preferences over sequences. But condition CI, implied by the discounted utility model, excludes complementarity of health outcomes over time. Therefore, sequence effects cannot be incorporated in the model.

Loewenstein (1987) and Loewenstein and Prelec (1991) present empirical evidence on savouring and dread as factors influencing individual intertemporal preferences. Savouring refers to the utility experienced through the anticipation of future pleasures such as better health. Dread refers to the disutility experienced through the anticipation of future unattractive events such as poor health. Savouring and dread are excluded in the discounted utility model by the assumption that an individual always prefers to receive positive health benefits sooner rather than later (impatience).

Adaptation, loss-aversion, savouring and dread challenge the discounted utility model also in another way. Their existence makes the isolation of the pure time preference effect very complicated. Analyses based on the discounted utility model need to isolate the pure time preference effect in order to determine the appropriate discount rate. However, it seems impossible to disentangle the separate influences of

[4]See for example Loewenstein (1987), Loewenstein and Thaler (1989), Loewenstein and Prelec (1991, 1992, 1993) and Frank and Hutchens (1993).

quantity, uncertainty, pure time preference, adaptation, loss-aversion, savouring and dread on individual intertemporal preferences. The discussion about the pure rate of time preference resembles the discussion about the concept of intrinsic risk attitude (Schoemaker, 1993): the concept is interesting as a theoretical construct, but unobservable in reality.

The confounding of various influences on intertemporal preferences can explain the anomalous preference patterns that have been observed with respect to health. For example, Redelmeier and Heller (1993) observe that a large proportion of their study population effectively applies a negative discount rate but that does not necessarily imply a negative pure rate of time preference. It can be explained by other influences, because the design used by Redelmeier and Heller (1993) is not capable of isolating the pure rate of time preference. Perhaps because health is a good with a time dimension inextricably bound to it, no experimental study can change the timing of the event without changing other factors, so attempts to measure the pure rate of time preference are likely to prove futile.

7.5.2. Direct evidence

Studies that have investigated the predictions of the discounted utility model with respect to health decision making have typically rejected the model. Lipscomb (1989) studied preferences over health streams by means of both the discounted utility model and a more general strategy (i.e., imposing less restrictions on the individual preference relation) which he refers to as the scenario strategy. Lipscomb observed some conflicting predictions, in the sense that the scenario strategy predicted a preference for health profile A over B where the discounted utility model predicted a preference for B over A. Since Lipscomb's scenario strategy imposes fewer restrictions, it will in general better predict choices, and in the case of conflicting predictions the discounted utility model seems to lead to the wrong prediction.

The results of recent empirical studies, attempting to elicit the rate of discount individuals apply to health outcomes, cast further doubts on the validity of the discounted utility model in modelling individual intertemporal preferences for health outcomes. The studies by Redelmeier and Heller (1993), Olsen (1993a), MacKeigan et al. (1993) and Cairns (1994) all reject the constant rate discounted utility model for the time preferences of such diverse groups as students, physicians, health policy makers and members of the general public. The pattern that emerges from these studies is a high discount rate for more proximate years and a lower discount rate for more distant years.

7.6 Variable rate discounted utility models

Given the deficiencies of the constant rate discounted utility model, Olsen (1993) and Harvey (1994) among others, have suggested replacing the constant rate discounted utility model by a variable rate discounted utility model. The axiomatization of the variable rate discounted utility model follows readily from the analyses of sections 2 and 3. In the context of section 2, condition CCI in addition to the structural assumptions is sufficient to obtain a general variable rate discounted utility model. In the context of section 3, additive independence has to be imposed as well. Variable rate discounted utility models are more general than the constant rate discounted utility model. Stationarity is no longer imposed, and therefore intertemporal trade-offs no longer need to be invariant with respect to the passage of time. Impatience does not need to be imposed, unless the discount function is to be a decreasing function of time.

Because variable rate discounted utility models make fewer assumptions with respect to individual preferences, they are better able to predict observable data. However, such required conditions as additive independence are still strong as has been argued in sections 4 and 5. Since CCI implies CI, sequence effects are still excluded. Finally, like its constant rate counterpart, the variable rate discounted utility model needs information on the pure rate of time preference, information that may be difficult to retrieve.

An individual whose preferences satisfy a variable rate discounted utility model at any point in time faces another problem: varying discount rates may lead to dynamically inconsistent preferences [Strotz, 1956; Hammond, 1976]. Suppose an individual must choose between two scenarios both involving three periods. Scenario A yields the sequence of health benefits $(0.8, 0.6, 0.4)$, and scenario B yields the sequence $(0.8, 0.4, 0.61)$. Suppose that the individual is a variable rate discounted utility maximizer at any time period. Assume that the discount rate for the first period is 0%, that the discount rate for the second period is 10% and for the third period 4%. It is easily checked that a variable rate discounted utility maximizer will prefer scenario B.

Suppose the individual reconsiders his choice after the first period. Suppose further that the individual can switch programs at a certain cost. Since benefits in the first period are equal in the two scenarios, the individual can concentrate on the future benefits of the two programs. Recalculating his discounted utility, the individual does not discount the benefits occurring in (what was) the second period and applies the discount rate of 10% to the benefits occurring in (what was) the third period. The individual will now prefer scenario A and will pay any amount up to the sum of money which is equivalent to the utility difference between the two programs

to be able to switch. Similar examples involving more than three periods can be constructed, in which the individual will pay an amount of money every period to be able to switch scenarios, only to end up in the scenario he already preferred in the first period, but not after having lost a good deal of money.

7.7 Concluding remarks

We have argued that the discounted utility model is inappropriate in modelling individual intertemporal preferences, both for certain health outcomes and for uncertain health outcomes. First, the axiom system of the discounted utility model is restrictive. Second, information about the individual pure rates of time preference for health, which is necessary for the discounted utility model, is unlikely to be retrievable. There is an additional problem in discounting health outcomes: the possibility of double discounting [Krahn and Gafni, 1993; Gafni, 1995]. Because health outcomes cannot be defined without reference to time duration, and utility assessment procedures typically introduce a time dimension [Torrance, 1986; Torrance and Feeny, 1989], individuals may incorporate their time preferences at least to some extent in the assessment of the utility of various health outcomes. Fully discounting health outcomes in such a situation would not be appropriate.

Where do these negative conclusions lead? One possibility is to relax the preference conditions underlying the discounted utility model to take individual intertemporal preferences for health outcomes into account in a more realistic way in health care program evaluation. On the other hand, relaxing preference conditions necessarily implies assessing more parameters. At every stage, the trade off between theoretical soundness and practical feasibility has to be made.

The variable rate discounted utility model does relax the preference conditions of the constant rate discounted utility model. However, the variable rate discounted utility model still does not allow complementarity between time periods, which is possibly the most restrictive assumption of the constant rate discounted utility model. Complementarity between times can be introduced in the model by following one of two approaches. One possibility is to extend the utility function by incorporating factors like habit formation [Pollak, 1970; Constantinides, 1990], the rate of benefit change [Frank and Hutchens, 1993] or preference/aversion for utility variation between adjacent periods [Gilboa, 1989]. Gilboa's model is an attempt to apply the Choquet expected utility models, which have been successful in decision making under uncertainty, to the time context. In the model where preferences concern risky health outcomes, preferences for health outcomes can be made to depend on the joint probability distribution, albeit in a limited sense, by relaxing additive independence to

mutual utility independence. Miyamoto and Eraker (1988) have found that utility independence generally holds in the health context.

Alternatively, the utility of health scenarios can be assessed directly by evaluating the whole stream of health outcomes. By not evaluating health outcomes separately, this approach in fact rejects coordinate independence. This is the idea behind Lipscomb's scenario strategy as well as the HYE [Gafni, 1995]. A disadvantage of scenario-based measures is that their "refusal" to evaluate health outcomes separately excludes the evaluation of health scenarios by short cuts. Whereas approaches based on coordinate independence need only assess a limited number of health states, in a scenario strategy every scenario must be assessed separately. This might limit their applicability in complex medical decision problems involving many possible health outcomes.

Appendix 1: Proof of theorem 7.1

Mathematical structure

The set of alternatives, X, is assumed to be a Cartesian product of the identical one period sets A: $X = A^T$. Time periods i are elements of a finite index set $I = \{1,\ldots,T\}$ with $T \in \mathbb{N}$, $\mathbb{N}$ denoting the set of natural numbers. The following structural assumptions are made with respect to A and X: (i) A is a connected and separable topological space[5]; (ii) X is endowed with the product topology. Connectedness ensures that every continuous function from X to $\mathbb{R}$, where $\mathbb{R}$ denotes the set of real numbers, has an interval as its image, so that this image has no holes. The weak order $\succsim$ defined on X is taken as primitive. A weak order $\succsim$ is complete ($x \succsim y$ *or* $y \succsim x$ for all $x,y \in X$) and transitive (if $x \succsim y$ and $y \succsim z$ then $x \succsim z$). This implies that the indifference relation $\sim$, defined as both $x \succsim y$ and $y \succsim x$, is an equivalence, i.e. it is symmetric ($x \sim y \Leftrightarrow y \sim x$), reflexive ($x \sim x$) and transitive). Strict preference $x \succ y$ is defined as $x \succsim y$ and not $y \succsim x$. We assume that $\succsim$ is continuous: $\{x : x \succsim y\}$ and $\{x : x \precsim y\}$ are closed for all $y \in X$. Continuity of the preference relation ensures that, if a function W exists that represents the preference relation, i.e. $W: X \rightarrow \mathbb{R}$ satisfies $x \succsim y \Leftrightarrow W(x) \geq W(y)$, then this function makes no jumps. The topological assumptions and the assumption that $\succsim$ is a continuous weak order are necessary and sufficient for the existence of a continuous representing function $W: X \rightarrow \mathbb{R}$ [Debreu, 1954].

We assume that there are at least two time periods and that every time period is essential, i.e. $x_{\cdot i}\alpha \succsim x_{\cdot i}\beta$ for some health outcomes α and $\beta \in A$ and for all i.

Proof of theorem 7.1

That (i) implies (ii) is straightforward. Hence we assume that (ii) holds and derive (i).

Definition A1.1: The preference order $\succsim$ is called persistent if
$(x_{\cdot i}\alpha) \succsim (x_{\cdot i}\beta) \Leftrightarrow (y_{\cdot j}\alpha) \succsim (y_{\cdot j}\beta)$ *for all* x,y,α,β,i,j.

---

[5] In fact topological separability does not have to be assumed if more than one time period is essential [Krantz, Luce, Suppes and Tversky, 1971 (section 6.11.1); Vind, 1990; Wakker, 1989 (Remark III.7.1)].

Persistence of the preference order asserts that preferences for health outcomes are identical in every time period. Persistence excludes to some extent a preference for variety. This can be seen for example by setting all elements of y in the above definition equal to α.

Persistence is implied by CCI. Set $x=y$, $v=w$, $\alpha=\beta$. By reflexivity of $\sim$: $x._i \alpha \sim x._i \alpha$ and $v._j \alpha \sim v._j \alpha$. So $x._i \alpha \preceq x._i \alpha$ and $v._j \alpha \succeq v._j \alpha$ both hold. Now $v._j \gamma \succeq v._j \delta$ follows from CCI.

By the structural assumptions being made, by CCI and by lemma IV.2.5 in Wakker (1989), we know from theorem IV.2.7 in Wakker (1989) that the preference relation can be represented by $x \succeq y \Leftrightarrow \sum \lambda_i V(x_i) \geq \sum \lambda_i V(y_i)$ with the λ_i uniquely determined and V continuous and unique up to positive linear transformations.

As shown above, CCI implies persistence. By persistence we cannot have $x._i \alpha \succeq x._i \beta \; \& \; x._j \alpha \prec x._j \beta$. Thus we cannot simultaneously have

$$\sum_{k \neq i} \lambda_k V(x_k) + \lambda_i V(\alpha) \geq \sum_{k \neq i} \lambda_k V(x_k) + \lambda_i V(\beta)$$

&

$$\sum_{l \neq j} \lambda_l V(x_l) + \lambda_j V(\alpha) < \sum_{l \neq j} \lambda_l V(x_l) + \lambda_j V(\beta)$$

From this it follows that either all λ_j are positive or all λ_j are negative. If all λ_j are negative, replace V by $-V$ and λ_j by λ_j. So all λ_j are positive. Then it automatically follows that if $\alpha_c \geq \beta_c$ then $V(\alpha) \geq V(\beta)$.

By impatience if $\alpha_c \geq \beta_c$ then $x._{i,i+1} \alpha, \beta \; \succeq \; x._{i,i+1} \beta, \alpha$. So

$$\sum_{k \neq i,i+1} \lambda_k V(x_k) + \lambda_i V(\alpha) + \lambda_{i+1} V(\beta) \geq \sum_{k \neq i,i+1} \lambda_k V(x_k) + \lambda_i V(\beta) + \lambda_{i+1} V(\alpha)$$

$$\Rightarrow \lambda_i [V(\alpha) \cdot V(\beta)] \geq \lambda_{i+1} [V(\alpha) \cdot V(\beta)]$$

Thus $\lambda_i \geq \lambda_{i+1}$. Set $\lambda_1 = 1$.

Now by stationarity

$(z._i \beta) \sim (z._{i+1} \alpha)$ *for all* $i, i+1$.

The existence of such a z follows from restricted solvability, which by lemma III.3.3 in Wakker (1989) is implied by the topological assumptions and by $\succsim$ being a continuous weak order. Restricted solvability asserts that for every $x_{\cdot i}\alpha \succsim y \succsim x_{\cdot i}\gamma$ there exists β such that $x_{\cdot i}\beta \sim y$. Take $y = z_{\cdot i+1}\alpha$, $z = x$ and select some γ_c such that $\alpha_c \succsim \beta_c \succsim \gamma_c$..

Because V is unique up to positive linear transformations, we can set $V(z)$ equal to zero. Following from the assumption that all time periods are essential, there exist α and β such that $V(\alpha), V(\beta) > 0$. Now from stationarity $(z_{\cdot 1}\beta) \sim z_{\cdot 2}\alpha) \Rightarrow V(\beta) = \lambda_2 V(\alpha) \Rightarrow \lambda_2 = V(\beta)/V(\alpha)$. Apply stationarity again to get $(z_{\cdot 2}\beta) \sim z_{\cdot 3}\alpha) \Rightarrow \lambda_2 V(\beta) = \lambda_3 V(\alpha) \Rightarrow \lambda_3 = \lambda_2[V(\beta)/V(\alpha)] \Rightarrow \lambda_3 = [V(\beta)/V(\alpha)]^2$.

Set $\pi = [V(\beta)/V(\alpha)]$. Then the constant discount rate model follows. Since by impatience $\pi \leq \lambda_1 = 1$ and every $\lambda_j > 0$ as established above, $0 < \pi \leq 1$.

Appendix 2: Proof of theorem 7.2

Mathematical structure

Z is defined as the set of all simple probability measures on X. A simple probability measure on X is a real-valued function P defined on the set of all subsets of X such that: *(i)* $P(B) \geq 0$ for every $B \subseteq X$; *(ii)* $P(X) = 1$; (iii) $P(B \cup C) = P(B) + P(C)$ when $B, C \subseteq X$ and $B \cap C = \varnothing$; *(iv)* $P(B) = 1$ for some finite $B \subseteq X$. A typical element of Z is denoted by $(p^1, x^1; \ldots; p^m, x^m)$ where, for each j, alternative x^j results with probability p^j and m can be any natural number. Elements of Z are denoted by capital Roman characters P, Q etc. Since Z contains many simple probability distributions, risky health outcomes can be mixed, or more formal, Z is closed under convex combinations: if $P, Q \in Z$ and $\lambda \in [0,1]$ then $\lambda P + (1-\lambda)Q \in Z$, where $\lambda P + (1-\lambda)Q$ is the lottery $(\lambda p^1 + (1-\lambda)q^1, x^1; \ldots; \lambda p^m + (1-\lambda)q^m, x^m)$.

The preference relation $\succsim_z$ is defined on Z. $\succsim_z$ is assumed to be a weak order. Furthermore, we impose the following two axioms [Jensen, 1967]:

1. vNM independence: $(P \succsim_z Q,\ 0 < \mu < 1) \Leftrightarrow (\mu P + (1-\mu)R) \succsim_z (\mu Q + (1-\mu)R)$ *for all* $P, Q, R \in Z$
2. Jensen continuity: $(P \succ_z Q \succ_z R) \Rightarrow \mu P + (1-\mu)R \succ_z Q$ *and* $Q \succ_z \kappa P + (1-\kappa)R$ *for some* $\mu, \kappa \in (0,1)$.

vNM independence is widely regarded to be the core of expected utility theory. It says that if P is weakly preferred to Q then any convex combination of P and R

should be weakly preferred to a similar convex combination of Q and R. Jensen continuity is an Archimedian condition, which asserts that, for $P \succ_z Q \succ_z R$, there are values μ and κ such that the convex combination of P and R is preferred to Q respectively Q is preferred to the convex combination of P and R.

Define Y as the set of simple probability distributions on the one period sets of health outcomes A. A marginal probability measure P_i is defined on A as: if $B \subset A$, then $P_i(B) = P(X : x_i \in B)$. A preference relation $\succeq$ is defined on Y from $\succeq$ on Z in the following way:

> $R \succeq_z S \Leftrightarrow P \succeq_z Q$ for $P,Q \in Z$ and $R,S \in Y$, such that $P_i = R$ and $Q_i = S$ for all time points i and P assigns probability one to a constant x.

Proof of theorem 7.2

That (i) implies (ii) is again straightforward. Hence, assume (ii). To derive is (i).

Because both $W = \sum \pi^{i-1} V(x_i)$ and U represent a preference relation $\succeq$ over X, they are related by a strictly increasing transformation. Under the assumptions of section 2, W is a continuous additive representation of $\succeq$ over X. Now, if U can also be written as a continuous additive representation of $\succeq$, then, by cardinality of V and U, U is a linear transform of V and can be taken identical to V. By additive independence $U(x) = \sum_i U_i(x_i)$ [this result has been proved by Fishburn (1965)]. By theorem 3.2 in Maas and Wakker (1994), U is continuous. Thus, U can be set equal to W: $U(x) = \sum_i V_i(x_i)$. Then apply the proof of theorem 7.1. This gives the desired result.

8

An empirical test of stationarity versus generalized stationarity[1]

Summary

This chapter presents an experimental test of the key axiom of the constant rate discounted utility model: stationarity. The results from the experiment display systematic violations of stationarity. The violations are in line with a phenomenon that Loewenstein and Prelec (1992) refer to as the "common difference effect." The presence of the common difference effect had been shown before for monetary outcomes. The results of this chapter suggest that it is also present in intertemporal choices involving health outcomes.

8.1 Introduction

In health economics, like in most areas of economics, the most common procedure to model the impact of differences in the temporal realization of outcomes is by applying a constant rate discounted utility model. According to the constant rate discounted utility model, the utility of a time stream of health outcomes $q_1,\ldots,q_T$ is evaluated by the following formula:[2]

$$U(q_1,\ldots,q_T)=\sum_{t=1}^{T}\beta^{t-1}U(q_t) \qquad (1)$$

where $U(q_t)$ is a utility function over health outcomes and β is a constant discount rate. Axiomatizations of this model can be found among others in Koopmans (1960,

[1] Based on Bleichrodt, H. and M. Johannesson, "Discounted utility models in health: An experimental test of stationarity versus generalized stationarity" (submitted for publication).

[2] This formula applies when time is discrete. The formula for the case where time is continuous is

$$\int_{t=0}^{T}U(q_t)e^{-\beta t}dt$$

1972), Fishburn (1970) and Fishburn and Rubinstein (1982). These authors acknowledge the fact that the underlying axioms of the constant rate discounted utility model are fairly restrictive and cannot be expected to hold in every decision context. Over the last decade, mainly by the work of Loewenstein (1987, 1988) and Loewenstein and Prelec (1992), the constant rate discounted utility model has been increasingly challenged as a description of individual intertemporal preferences for monetary outcomes. Various anomalies have been identified and alternative theories have been proposed. Empirical evidence further showed that violations of the axioms underlying the constant rate discounted utility model are easily obtained and do not require ingenious experimental designs to be revealed. In the context of health decision making, Bleichrodt and Gafni (1995) have provided arguments why the constant rate discounted utility model may fail. Empirical studies examining individual intertemporal preferences for health outcomes typically reject the predictions of the constant rate discounted utility model.[3]

The aim of this chapter is to present an empirical test of the appropriateness of the constant rate discounted utility model in health. Our study differs from earlier studies in health in that we test the key axiom of the constant rate discounted utility model rather than the predictions of the complete model. The key axiom of the constant rate discounted utility model is a condition that is referred to in the literature as stationarity. Alternative intertemporal models have focused on generalizing stationarity retaining the other conditions on which the constant rate discounted utility model relies. The motivation to generalize stationarity was an empirically observed characteristic of intertemporal choice to which Loewenstein and Prelec refer as "the common difference effect." This chapter describes an experiment in which the null hypothesis corresponding to stationarity is tested against the alternative hypothesis corresponding to the common difference effect. This design allows us to draw inferences with respect to the contribution that models based on generalized stationarity can make to the explanation of individual intertemporal preferences for health.

The structure of the chapter is as follows. In section 2 we discuss in more detail the contents of stationarity, generalized stationarity and the common difference effect. This discussion allows the formulation of empirical tests of stationarity and generalized stationarity. The experiment designed to perform these empirical tests is described in section 3. Section 4 discusses the methods used to analyze the experimental data. Section 5 presents results. Section 6 contains concluding remarks. The appendix contains an algebraic derivation of a claim made in section 3.

[3]Cf. Olsen (1993), Cairns (1994), Redelmeier and Heller (1993), Mackeigan et al. (1993).

8.2 Stationarity and generalized stationarity

Let q_t denote a quality of life level occurring at time period t. Denote a health profile, by which we mean a temporal sequence of health outcomes, by $(q_1,\ldots,q_T)$. We are interested in a preference relation $\succeq$ over health profiles, meaning "at least as preferred as." Throughout $\succeq$ is assumed to be complete and transitive. The asymmetric part of $\succeq$ (strict preference) is denoted by $\succ$, and the symmetric part of $\succeq$ (indifference) by $\sim$. A function U is said to represent a preference relation $\succeq$ if and only if $(q_1,q_2,\ldots,q_T) \succeq (q'_1,q'_2,\ldots,q'_T)$ implies $U(q_1,q_2,\ldots,q_T) \geq U(q'_1,q'_2,\ldots,q'_T)$.

The preference relation $\succeq$ is assumed to satisfy certain conditions[4] such that it can be represented by the utility function

$$U(q_1,\ldots,q_T)=\sum_{t=1}^{T} \lambda_t U(q_t) \qquad (2)$$

where λ_t is a positive, period-specific scaling factor, which can be interpreted as a discount factor. The λ_t are generally assumed to be monotonically decreasing over time. Bleichrodt and Gafni (1995), among others, have shown that monotonically decreasing λ_t correspond to "impatience", i.e. individuals prefer favourable outcomes to occur sooner rather than later. The representation given in equation (2) underlies all models to be discussed in this chapter.

The difference between the models we discuss lies in the preference condition that is imposed on the discount factors λ_t. The characterizing condition of the constant rate discounted utility model is *stationarity*. In the formulation by Fishburn (1970), stationarity imposes the following restriction on the intertemporal preference relation:

$$(q_1,\ldots,q_{T-1},q^c) \succeq (q'_1,\ldots,q'_{T-1},q^c)$$
$$\text{if and only if } (q^c,q_1,\ldots,q_{T-1}) \succeq (q^c,q'_1,\ldots,q'_{T-1}) \qquad (3)$$

for some health outcome q^c, common to both vectors. In words, stationarity says that the preference relation should be invariant if each health outcome is advanced by one period and q^c is shifted from the last period to the first. It is straightforward to show that if stationarity holds for *some* q^c, it will hold for all q^c, given that the preference relation can be represented by equation (2).

---

[4] Cf. e.g. Wakker (1984).

The impact of stationarity is to make preferences invariant with respect to the passage of time. This is most easily seen for health profiles that differ only at two points in time. Denote by $z_{.t}q$ the health profile in which the constant health outcome z occurs at all points in time except point in time t at which health outcome q occurs. Consider preferences over two health profiles $z_{.r}q$ and $z_{.s}q'$. For convenience and without loss of generality we assume that both q and q' are stricly preferred to z Suppose $z_{.r}q \succsim z_{.s}q'$. Then by stationarity $z_{.(r+t)}q \succsim z_{.(s+t)}q'$. Repeatedly applying stationarity gives $z_{.(r+e)}q \succsim z_{.(s+e)}q'$ for any e. Thus, by stationarity preferences between two health profiles depend only on the difference in time of realization between distinguishing health outcomes and not on the exact point in time at which these distinguishing health outcomes are realized. That is, preferences between health profiles do not depend on the passage of time. The above argument can easily be generalized to the case where the health profiles differ at more than two points in time.

Experimental tests of stationarity, involving monetary outcomes, have rejected stationarity. Loewenstein and Prelec (1991) observed that individual intertemporal preferences follow a pattern they refer to as the common difference effect. The common difference effect predicts that for two points in time r and s, with r strictly preceding s $(r<s)$:

$$if\ z_{.r}q \sim z_{.s}q'\ then\ z_{.(r+e)}q \prec z_{.(s+e)}q'\ for\ \ z \prec q \prec q';\ r<s;\ e>0 \qquad (4)$$

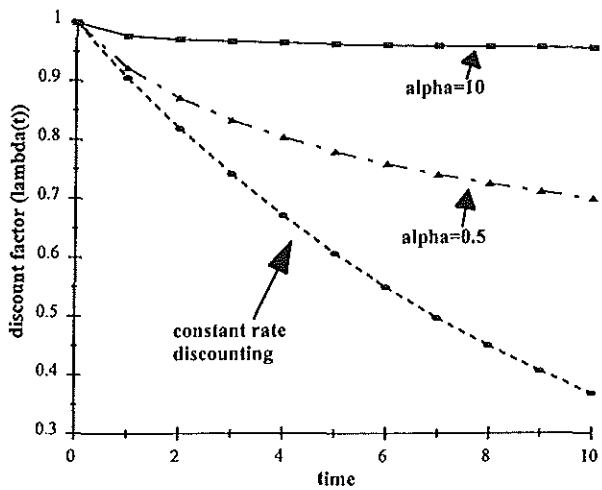
In words, the common difference effect asserts that constant differences in timing between two outcomes loom larger the less remote they are in time. In equation (4), the difference in timing between the outcomes is in both preference comparisons equal to s-r. In the former preference comparison, which is less remote in time, indifference holds (i.e., the difference in timing is sufficient to offset the difference in utility between q and q'). In the second preference comparison the second profile is strictly preferred (i.e., the difference in timing is not sufficient to offset the difference in utility between q and q'). To account for the common difference effect, Loewenstein and Prelec (1992) propose to generalize stationarity to the following condition:

$$if\ z_{.r}q \sim z_{.s}q'\ then\ \ \exists k(q,q') \geq 1\ such\ that\ \ z_{.[r+e]}\,q \sim z_{.[ku+s]}\,q'$$
$$for\ z \prec q \prec q';\ r<s \qquad (5)$$

Note that k depends on q and q'. If k is greater than 1, for all $q \prec q'$, then this generalized stationarity principle allows the common difference effect. In that case the difference in timing has to increase to restore indifference when the outcomes are

more remote in time. If $k=1$ for all $q \prec q'$ equation (5) reduces to equation (2). Thus, the stationarity principle used in the characterization of the the constant rate discounted utility model is a special case of this generalized stationarity condition.

Figure 8.1: The generalized discount function for some values of α



Loewenstein and Prelec (1992) show that imposing generalized stationarity with $k \geq 1$ on the representation (2) leads to a discount function of the following form:

$$\lambda(t)=(1+\alpha t)^{-\beta/\alpha}; \quad \alpha,\beta > 0 \qquad (6)$$

The limiting case where $\alpha \downarrow 0$ corresponds to the constant rate discounted utility model. The parameter α reflects the importance attached to future periods. Holding β constant, the higher α is, the more weight is given to future periods. This can for example be seen in figure 8.1 which shows the discount function $\phi(t)$ for various values of α holding β fixed at 0.10.

Because $\alpha > 0$, the constant rate discounted utility model is the limiting case in which the future receives the lowest weight. In the sequel of this chapter, when we speak about generalized stationarity and generalized discounted utility models we mean the situation where $k > 1$. That is, by generalized discounted utility models we mean models that satisfy the common difference effect.

8.3 Experimental design

8.3.1 Subjects and health states

Eighty students at the Stockholm School of Economics and ninety-two students at Erasmus University Rotterdam took part in the experiment. The students were paid about $15 in local currency for their participation. The experiment was carried out in 17 sessions lasting approximately one hour with on average ten respondents per session. The procedure followed in each session was to explain first a specific task to respondents, then to ask respondents to perform the specific task and then to explain the next task. A "master" version of the experiment was designed in English. This "master" version was subsequently translated into Swedish and Dutch. Before drafting the final version, we tested the questionnaire extensively both in Stockholm and in Rotterdam using faculty staff members as respondents.

Table 8.1: Health states A and B

A	B
1. able to perform all tasks at home and/or at work without difficulties	1. able to perform all tasks at home and/or at work albeit with some difficulties
2. able to perform all self care activities (eating, washing, dressing) without help	2. able to perform all self care activities (eating, washing, dressing) without help
3. able to participate in all types of leisure activities albeit with some difficulties	3. unable to participate in certain types of leisure activities
4. now and then light to moderate pain and/or other complaints	4. often light to moderate pain and/or other complaints

We selected eight health states to be included in the questionnaire. The health states were taken from the Maastricht Utility Measurement Questionnaire, a slightly adapted version of the McMaster Health Utility Index [Bakker et al., 1995; Rutten-van Mölken et al., 1995]. The selected health states correspond to commonly occurring types of back pain and rheumatism. Health states in the Maastricht Utility Measurement Questionnaire consist of six dimensions. We excluded two dimensions from the health state descriptions: side effects of medicines and anxiety about prognosis. These dimensions were excluded because they were not essential

for the purpose of this experiment and because giving too much information would unnecessarily complicate the tasks respondents were faced with. Thus health states consisted of four dimensions: general daily acivities, self care, leisure activities, and pain. The health states were indicated by capital letters and were described on a set of cards, which were handed out to respondents at the beginning of each session. Health states A and B are relevant for the analysis of this chapter. They are described in table 8.1.

8.3.2 Empirical test

The questionnaire was divided into six sections. Sections 1-5 were aimed to test hypotheses that are not relevant for the purpose of the present chapter.

Section 6 was designed to test stationarity against generalized stationarity. As has been outlined in section 2, the critical factor that distinguishes stationarity from generalized stationarity is the impact of the remoteness in time of the outcomes. By stationarity, intertemporal preferences are invariant with respect to remoteness in time of the outcomes as long as the difference between the times at which the outcomes are realized (the realization times) is held constant. By generalized stationarity, intertemporal preferences do not only depend on the difference in time of realization, but also on the remoteness in time: the more remote the outcomes are in time the less important the difference in time of realization becomes. That is, to retain indifference between profiles, the more remote the outcomes are in time, the greater the difference in time of realization between the outcomes must be. Formally, by the common difference effect: if $z_{.,r}q \sim z_{.,s}q'$ then $z_{.,r+e}q \prec z_{.,s+e}q'$ $(e > 0)$. Given impatience $z_{.,r+e}q \sim z_{.,s+f}q'$, $e < f$. Thus, the difference in time of realization between the outcomes has to increase if indifference is to be retained. The impact of remoteness in time on the difference in realization time is the hypothesis that we originally set out to test. It turned out in the pilot sessions that respondents had problems to compare profiles in which the realization time varied between the two options. We therefore opted for a different set up in which the realization time was equal between the two options.

Consider a choice between two profiles. The first profile consists of y periods in health state q' preceded and followed by health state z, which we take for convenience equal to full health. The second profile consists of $x < y$ periods in health state $q \prec q'$, also preceded and followed by full health. The total life span is T. When the time at which health state q is realized is r, we can write these two profiles as $z_{.,r,\ldots,r+y}q'$ and $z_{.,r,\ldots,r+x}q$ respectively. Suppose that for a particular realization time r, $z_{.,r,\ldots,r+y}\, q' \sim z_{.,r,\ldots,r+x}\, q$, i.e. an individual is indifferent between the two

profiles. Then it follows from the argument of section 2 that by repeated application of stationarity the individual will also be indifferent between the two profiles for any realization time $r' \neq r$. That is, by stationarity $z_{.r,\ldots,r+y}q' \sim z_{.r,\ldots,r+x}q \Leftrightarrow z_{.r',\ldots,r'+y}$ $q' \sim z_{.r',\ldots,r'+x}q$. In the appendix we show that generalized discounted utility models predict that if $z_{.r,\ldots,r+y}q' \sim z_{.r,\ldots,r+x}q$ then for $r<r'$, $z_{.r',\ldots,r'+y}$ $q \prec z_{.r',\ldots,r'+x}q$. Given that health state q is less preferred than full health and given the model represented by equation (2), increases in the time in q, x, will decrease the attractiveness of the second profile. If x is increased sufficiently, indifference between the two profiles will be restored. In the experiment we aimed to test the impact of variations in realization time on the time in q, x. Our null hypothesis, which corresponds to stationarity, is that x will remain constant with changes in the realization time. Our alternative hypothesis, which corresponds to generalized stationarity is that x will increase with increases in realization time.

8.3.3 Stimuli

We asked respondents to make a choice between three pairs of health profiles. All profiles had a life duration of 20 years ($T=20$). The first option always consisted of four years in health state A ($y=4$; $q'=A$) followed by full health. The time at which the individual's health fell to health state A, the realization time r, varied across the three questions. In the first question this happened immediately ($r=0$), in the second question this happened after one year ($r=1$) and in the third question this happened after three years ($r=3$). The second option differed from the first in that the individual's health fell to health state B ($q=B$) for a specified time duration x rather than to health state A for four years. After x years in B the individual's health was restored to full health. In section 1 of the questionnaire, respondents had ranked health states relative to full health. In the pilot session all respondents ranked full health above health state A which in turn was ranked above health state B. Therefore our choice of health states seemed appropriate. This was confirmed in the actual experiment: all respondents indicated the rank ordering *full health* $\succ$ *health state A* $\succ$ *health state B*. We used five different time durations x for the time spent in B: 1 year, 1.5 year, 2 years, 2.5 years and 3 years. The time duration x varied across the three questions and was determined by a random process (by draws from the standard normal distribution). Questionnaires were distributed randomly across respondents. We made sure that an equal number of observations was obtained on each time duration. Table 8.2 gives an example of a combination of the three questions that was used in the experiment.

Table 8.2: Example of the three questions asked

	Question 1	Question 2	Question 3
Option 1	4 years in *A*, 16 years in full health	1 year in full health, 4 years in *A*, 15 years in full health	3 years in full health, 4 years in *A*, 13 years in full health
Option 2	1 year in *B*, 19 years in full health	1 year in full health, 2.5 years in *B*, 16.5 years in full health	3 years in full health, 1.5 years in *B*, 15.5 years in full healh

Given the nature of health states *A* and *B*, corresponding to different types of back problems, respondents in the pilot sessions indicated that they considered profiles in which *A* and *B* were temporary to be realistic.[5] We asked respondents to make a choice between the two options rather than to indicate a specific time in *B* for which they considered the two profiles to be be equivalent. Experience from the pilot sessions indicated that respondents found choice questions easier to deal with than matching questions.

A problem of asking the three questions consecutively might be that responses are biased by anchoring. By asking the three questions consecutively the similarity between the questions was emphasized and this may have induced respondents to follow the same strategy in all three questions even though the time in *B* varied per question and was determined randomly. We deliberately selected this type of format. If anchoring is present it would bias responses in the direction of stationarity. If respondents follow a similar strategy in all three questions then they will tend to give similar answers and thus the mean/median time in *B* will be equal across questions, which supports stationarity. We believed that if we would observe violations of stationarity, even though our experimental design may have favoured stationarity, this would count as strong evidence against the descriptive validity of stationarity.[6]

[5] At the end of the actual experiment we asked respondents whether they considered the choices realistic. Most respondents said they found the choices relatively easy and realistic.

[6] In other sections of the questionnaire we faced similar problems of anchoring. In these sections we explicitly tested for anchoring by including questions that were not susceptible to possible anchoring bias. We compared the responses to these questions with the responses to the same questions where these might have been susceptible to anchoring bias. This test indicated that anchoring was not a serious problem for these responses [cf. chapter 4].

8.4 Methods

The observed dependent variable, the choice between profile 1 and profile 2, is discrete. An appropriate estimator is therefore either the probit or the logit model. For univariate models, these models are almost indistinguishable [Greene, 1993]. The three models we estimated were all of the form

$$y_i^* = \beta_0 + \beta_1 * t_i + \varepsilon_i \qquad (11)$$

where y^* is a latent variable that can be interpreted as "inclination to choose option 2," t denotes the time spent in health state B and ε is an error term. We expected β_1 to be negative, given that the longer the time spent in B the lower this inclination will be. Estimation of binary choice models is usually based on the method of maximum likelihood. Yatchew and Griliches (1985) have shown that if the disturbances in the regression underlying the binary choice model are heteroskedastic, the maximum likelihood estimators are inconsistent and the variance matrix is inappropriate.

We included two tests of heteroskedasticity. The tests we used were based on a correction for heteroskedasticity of the form $var[\varepsilon_i] = exp[\gamma * t]$. Davidson and MacKinnon (1984) have argued that this test is not appropriate in the logit model. For this reason we decided to use the probit model. We used two asymptotically equivalent test statistics to test for heteroskedasticity: a Likelihood Ratio test and a Lagrange Multiplier test proposed by Davidson and MacKinnon, to which they refer as LM_2. Davidson and MacKinnon present Monte Carlo evidence that LM_2 has the best finite sample properties in comparison to several alternative, asymptotically equivalent, Lagrange Multiplier tests. Davidson and MacKinnon's Monte Carlo evidence further shows that LM_2 performs better under the null hypothesis of no heteroskedasticity than the Likelihood Ratio test. However, LM_2 has less power than the Likelihood Ratio test. In case the null hypothesis of no heteroskedasticity had to be rejected, we estimated the heteroskedasticity-adjusted model

$$y_i^* = [\beta_0 + \beta_1 * t_i] / [exp(\hat{\gamma} * t_i)] \qquad (12)$$

Once consistent estimates are obtained for β_0 and β_1, the mean time in B can, by a corollary to proposition 1 in Kriström (1990a), be calculated as $\hat{\beta}_0 / \hat{\beta}_1$. By the symmetry property of the standard normal distribution, the mean is equal to the median. By means of an approximation argument, the following formula can then be obtained for the variance of the mean (median) time in B [cf. e.g. Abdelbasit and Plackett (1983)]:

$$var(\hat{\beta}_0/\hat{\beta}_1) \cong (\hat{\beta}_0/\hat{\beta}_1)^2 [var(\hat{\beta}_0)/\hat{\beta}_0^2 + var(\hat{\beta}_1)/\hat{\beta}_1^2 - 2*cov(\hat{\beta}_0,\hat{\beta}_1)/(\hat{\beta}_0*\hat{\beta}_1) \quad (13)$$

The asymptotic covariance matrix for the maximum likelihood estimator can be estimated in one of three ways [Greene, 1993]. We used the negative inverse of the actual Hessian of the log likelihood.

Goodness of fit of the model was assessed both by individual prediction and by the Likelihood Ratio Index (alternatively called McFadden's pseudo R^2). Individual prediction is the percentage of binary responses correctly predicted by the model. The Likelihood Ratio Index (LRI) is calculated as $[1-(lnL/lnL_0)]$, where $ln\ L$ stands for the maximized value of the log likelihood function and $ln\ L_0$ stands for the maximized log likelihood computed with only a constant term, i.e. $\beta_1=0$. The LRI has intuitive appeal in that it is bounded by 0 and 1. However, it does not correspond to any of the R^2 measures in the linear regression model. Values between 0 and 1 have no natural interpretation.

A disadvantage of using the probit model is that the error terms in the equation for the latent variable are assumed to be normally distributed. If the assumption of normality does not hold the parameter estimates will not be consistent. We therefore also examined the data by means of a distribution-free estimator developed by Kriström (1990b). Let $\hat{\pi}_i$ be the observed proportion of respondents choosing the second option when the time in B is t_i. Our experiment produces a sequence of proportions $(\hat{\pi}_1,\hat{\pi}_2,\ldots,\hat{\pi}_m)$, where we use the convention that $\hat{\pi}_1$ corresponds to the lowest time in B. Kriström's method makes use of a theorem by Ayer et al. (1955) which shows that if the $\hat{\pi}_i$ form a monotone non-increasing sequence, then this sequence provides a distribution free maximum likelihood estimator of the true probability of choosing the second option. In case the sequence of observed proportions is increasing, an adjustment has to be made. This adjustment was not necessary in our study because the proportions formed a non-increasing sequence. In Kriström's method the proportion of respondents choosing the second option for each t_i is used to construct an empirical "survival" function with respect to delay time. This function is then integrated to obtain the mean time in B. The median time in B can be obtained by calculation of the time in B for *which P (choose second option) = 0.5*. In order to be able to integrate the empirical survival function we had to assume that $\hat{\pi}_i = 0$ for 4 years in B and that $\hat{\pi}_i = 1$ for 0 years in B, i.e. 20 years in full health. Both assumptions are entirely plausible in the context of our experiment.

8.5 Results

Table 8.3: Results of the probit estimation

	Question 1	Question 2	Question 3
Parameter estimates			
$\hat{\beta}_0$	1.764	2.454	2.224
(standard error)	(0.329)	(0.363)	(0.343)
$\hat{\beta}_1$	-0.996	-1.276	-1.082
(standard error)	(0.161)	(0.176)	(0.163)
Log-Likelihood	-94.739	-84.224	-93.170
Goodness of fit			
Likelihood Ratio Index	0.188	0.280	0.217
% correctly predicted	73.1	78.1	72.1
Heteroskedasticity			
LM_2 ($\chi^2(1)$)	$\chi^2(1) = 3.955$	$\chi^2(1) = 0.881$	$\chi^2(1) = 0.062$
(p-value)	($p = 0.046$)	($p = 0.348$)	($p = 0.807$)
Likelihood Ratio ($\chi^2(1)$)	$\chi^2(1) = 4.690$	$\chi^2(1) = 0.975$	$\chi^2(1) = 0.076$
(p-value)	($p = 0.030$)	($p = 0.323$)	($p = 0.786$)
Mean time in B (standard error)			
Without correction for heteroskedasticity	1.772 (0.109)	1.924 (0.088)	2.056 (0.098)
Corrected for heteroskedasticity	1.575 (0.083)	n.a.	n.a.

Table 8.3 displays the results of the probit estimation. The parameter estimates differ for the three questions. This is also reflected in the estimated mean times in *B* for which indifference holds between the profiles. There is a clear pattern in the mean times in *B* for which indifference holds: the later the time of realization, the longer the mean time in *B*. The observed pattern violates stationarity which, as has been observed above, predicts that the mean time in *B* should be equal across the questions. The pattern supports generalized stationarity and thus the common difference effect.

As we have derived in the appendix, generalized stationarity predicts that the mean time in *B* will be highest in question 1 and lowest in question 3. This is the pattern we observe. These findings suggest that respondents attach more weight to future outcomes than assumed by the constant rate discounted utility model. However, only the difference in time in *B* between questions 1 and 3 is statistically significant ($p < 0.01$). [7]

The two tests for heteroskedasticity both reject the null hypothesis of no heteroskedasticity for question 1. We therefore re-estimated this equation applying a correction for heteroskedasticity. Applying the tests to the re-estimated equation did not show any further evidence of heteroskedasticity. Both tests could not reject the null hypothesis of no heteroskedasticity for questions 2 and 3. Corrected for heteroskedasticity, the mean time in *B* decreased to 1.575 years in question 1. This makes the observed pattern even more apparent. Both the differences in mean time in *B* between question 1 and question 2 and between question 1 and question 3 are highly significant *($p < 0.001$)*

Table 8.4: Results of the non-parametric estimation

	Question 1	Question 2	Question 3
mean time in *B*	1.82	1.93	2.05
median time in *B*	1.54	1.78	2.10

Table 8.4 displays the results of the non-parametric estimation. The non-parametric analysis confirms the pattern that was revealed by the probit estimation. The results of our experiment suggest a violation of stationarity. The violation is in the direction predicted by generalized stationarity models satisfying the common difference effect. This holds regardless of whether the median or the mean is used to analyze the data. In the latter case the pattern is less pronounced.

[7] In testing for statistical significance we implicitly assume that responses are independent, i.e. that there is no anchoring. If there is anchoring, responses will be positively correlated and the standard deviation we used in testing for statistical significance will be too high. Thus, if responses are anchored statistical significance will be observed for smaller differences. Given this, our tests of significance can be considered conservative in that we are less likely to find significance of differences.

8.6 Concluding remarks

This chapter presents an empirical test of stationarity. Stationarity is the condition that distinguishes the constant rate discounted utility model within the class of intertemporally separable models. The results of our experiment display violations of stationarity. The violation is most pronounced when comparing the immediate future (i.e. the time before the earlier outcome is realized is zero) with the later future (the time before the earlier outcome is realized is greater than zero). This is consistent with an "immediacy effect," which has been shown to produce a virtual discontinuity of preference in intertemporal choice involving monetary outcomes and which is comparable to the certainty effect in choice under uncertainty [Prelec and Loewenstein, 1991].

The violations are in the direction predicted by the common difference effect. The results of our experiment suggest that the common difference effect, which has been shown to be present in intertemporal choices involving monetary outcomes, is also present when intertemporal choices involve health outcomes.

We have no reason to suspect that the violations are caused by the hypothetical character of the experiment. Respondents typically indicated that they found the choices relatively easy to make and that they considered the profiles realistic. If respondents would have had problems answering the questions, the most logical strategy for them to follow would have been to give similar answers to all three questions. Such a strategy would have supported stationarity. The fact that our results suggest violations of stationarity indicates that respondents have not followed a random strategy, but have made deliberate choices. This conclusion is further supported by the fact that the mean time in B was a significant variable in all three equations. The results of our experiment suggest that it is possible to ask individuals to make intertemporal trade-offs of the kind reported in this chapter. This is an important observation for future research.

Appendix: Derivation

In the appendix we derive the claim made in section 3 that a consequence of the generalized discounted utility model is that to maintain indifference between the profiles the time in q should increase with increases in the realization time r. The claim follows from straightforward, but somewhat tedious application of algebra.

Suppose that $z_{\cdot,r+1,\ldots,r+y}\,q' \sim z_{\cdot,r+1,\ldots,r+x}\,q;\ q \prec q',\ x<y$. In terms of the generalized discounted model this indifference implies that

$$U(H)*\sum_{t=0}^{t=r-1}(1+\alpha t)^{-\beta/\alpha}+U(A)*\sum_{t=r}^{t=r+y}(1+\alpha t)^{-\beta/\alpha}+U(H)*\sum_{t=r+y+1}^{t=T}(1+\alpha t)^{-\beta/\alpha}$$

=

$$U(H)*\sum_{t=0}^{t=r-1}(1+\alpha t)^{-\beta/\alpha}+U(B)*\sum_{t=r}^{t=r+x}(1+\alpha t)^{-\beta/\alpha}+U(H)*\sum_{t=r+x+1}^{t=T}(1+\alpha t)^{-\beta/\alpha} \qquad (A1)$$

After elimination of common terms we obtain

$$U(A)*\sum_{t=r}^{t=r+y}(1+\alpha t)^{-\beta/\alpha}=U(B)*\sum_{t=r}^{t=r+x}(1+\alpha t)^{-\beta/\alpha}+U(H)*\sum_{t=r+x+1}^{t=r+y}(1+\alpha t)^{-\beta/\alpha} \qquad (A2)$$

or

$$[U(A)-U(B)]=[U(H)-U(A)]*\{\sum_{t=r+x+1}^{t=r+y}(1+\alpha t)^{-\beta/\alpha}\}/\{\sum_{t=r}^{t=r+x}(1+\alpha t)^{-\beta/\alpha}\} \qquad (A3)$$

Now consider a different realization time $r'=r+e;\ e>0$. Suppose the time in q' is still y and the time in q still x. Then the utility of the first option, $U(z_{\cdot,r',\ldots,r'+y}\,q')$ is equal to

$$U(H)*\sum_{t=0}^{t=r'-1}(1+\alpha t)^{-\beta/\alpha}+U(A)*\sum_{t=r'}^{t=r'+y}(1+\alpha t)^{-\beta/\alpha}+$$
$$U(H)*\sum_{t=r'+y+1}^{t=T}(1+\alpha t)^{-\beta/\alpha} \qquad (A4)$$

and the utility of the second option, $U(z_{.r',\ldots,r'+x}q)$ is equal to

$$U(H)*\sum_{t=0}^{t=r'-1}(1+\alpha t)^{-\beta/\alpha}+U(B)*\sum_{t=r'}^{t=r'+x}(1+\alpha t)^{-\beta/\alpha}+$$
$$U(H)*\sum_{t=r'+x+1}^{t=T}(1+\alpha t)^{-\beta/\alpha} \qquad (A5)$$

For clarity of the subsequent argument it will be more convenient to denote time by t' when the summation signs apply to realization time r'. Eliminating common terms and rearranging leaves a comparison between

$$[U(A)-U(B)]*\sum_{t'=r'}^{t'=r'+x}(1+\alpha t)^{-\beta/\alpha} \quad and$$
$$[U(H)-U(A)]*\sum_{t'=r'+x+1}^{t'=r'+y}(1+\alpha t)^{-\beta/\alpha} \qquad (A6)$$

Substituting for $[\,U(A)-U(B)\,]$ from (A3) and rearranging terms leaves a comparison between

$$[U(H)-U(A)]*\sum_{t'=r'}^{t'=r'+x}(1+\alpha t)^{-\beta/\alpha}*\sum_{t=r+x+1}^{t=r+y}(1+\alpha t)^{-\beta/\alpha} \qquad (A7a)$$

and

$$[U(H)-U(A)]*\sum_{t'=r'+x+1}^{t'=r'+y}(1+\alpha t)^{-\beta/\alpha}*\sum_{t=r}^{t=r+x}(1+\alpha t)^{-\beta/\alpha} \qquad (A7b)$$

Eliminating the common positive constant $[U(H)-U(A)]$, substituting $r'=r+e$, and rearranging gives a comparison between

$$\sum_{t=r+x+1}^{t=r+y}\sum_{t'=r+e}^{t'=r+x+e}[(1+\alpha t)*(1+\alpha t')]^{-\beta/\alpha} \qquad (A8a)$$

and

$$\sum_{t=r}^{t=r+x}\sum_{t'=r+x+e+1}^{t'=r+y+e}[(1+\alpha t)*(1+\alpha t')]^{-\beta/\alpha} \qquad (A8b)$$

Select two arbitrary values for t, one from the interval $[r;r+x]$ and one from the interval $[r+x+1;r+y]$. Let the former value of t be denoted by l and the latter by m. Then corresponding elements of the sequences given in equations (A8a) and (A8b) are

$$[(1+\alpha m)^{*}(1+\alpha(l+k))]^{-\beta/\alpha} \quad and \quad [(1+\alpha l)^{*}(1+\alpha(m+k))]^{-\beta/\alpha} \qquad (A9)$$

Rewriting the terms between square brackets [ ] gives

$$[1+\alpha(l+k+m)+\alpha^{2}(l+k)m] \quad and \quad [1+\alpha(l+k+m)+\alpha^{2}(m+k)l] \qquad (A10)$$

Eliminating common terms leaves a comparison between $\alpha^{2}km$ and $\alpha^{2}kl$. Given that m is taken from the interval $[r+x+1;r+y]$ and l from the interval $[r;r+x]$, m is greater than l and thus the first term is greater than the second. However, given that both α and β are positive, we have to raise both terms to a negative power. Thus the total term is greater for the second expression. This finally establishes that the utility of the second option is greater than the utility of the first option. To restore indifference, the utility of the second option has to be decreased. This is ensured by increasing the time in B relative to the time in full health.

9

Health utility indices and equity considerations[1]

Summary

Concern has been expressed about the equity implications of QALY-based decision making. The aim of this chapter is to propose methods that incorporate equity concerns into cost utility analysis. Two interpretations of QALYs are considered: QALYs as (von Neumann Morgenstern) utilities and QALYs as measures of health. A justification is provided for aggregating "QALYs as utilities" over individuals. The conditions underlying the unweighted aggregation of QALYs over individuals are identified. Two types of equity algorithms are proposed by relaxing some of these conditions: algorithms that take into account the final distribution of QALYs (ex post equity) and an algorithm that takes into account both ex post equity and ex ante equity.

9.1 Introduction

Utility indices for health care programs, such as QALYs, have been criticized for being primarily concerned with efficiency, ignoring equity implications.[2] The importance of incorporating equity considerations into cost utility analysis has been widely acknowledged by researchers in the field [e.g. Williams, 1993]. However, despite statements of intent, few attempts have been made thus far to actually develop methods by means of which equity considerations can be taken into account in cost utility analysis. One of the few exceptions is Wagstaff (1991) in which it is suggested to combine equity and efficiency considerations in cost utility analysis by means of the social welfare function underlying Atkinson's (1970) index of inequality. However, Wagstaff did not pursue this idea any further, in particular

[1] Based on Bleichrodt, H., "Health utility indices and equity considerations," (submitted for publication).

[2] cf. e.g. Lockwood (1988); Harris (1988); Smith (1987); Broome (1988); Broome (1993).

he did not indicate how the parameters of this social welfare function can be assessed by experimental methods.

The aim of this chapter is to derive functional forms that allow trading off the efficiency gains of a health care program against its equity implications.[3] Given that QALYs are the most frequently used outcome measure in cost utility analysis, we will refer to the gains of a health care program as the number of QALYs gained. This does not restrict the generalizability of the analysis of the chapter. One might as well substitute other health utility indices in the algorithms to be derived. Regardless of which outcome measure is used, if societal decisions are to be based on individual values, decisions with respect to the aggregation of these individual values have to be made. The functional derivations presented in this chapter are based on the tools of multi-attribute utility theory. Multi-attribute utility theory has been developed as a procedure to make explicit the trade-off between conflicting objectives. Two interpretations of QALYs that have been distinguished in the literature are considered: QALYs as von Neumann Morgenstern (vNM) utilities and QALYs as measures of health.[4] It has been claimed that the QALYs as utilities approach lacks a theoretical foundation, since utilities cannot be interpersonally compared in a meaningful way. We will address this problem in sections 3 and 4.

The structure of the chapter is as follows. Section 2 discusses the two interpretations of QALYs. Sections 3 and 4 provide a rationale for aggregating "QALYs as utilities" over individuals. In section 3 it is argued that if we want to incorporate equity considerations in cost utility analysis, full interpersonal comparability of utilities is needed. In section 4 an argument is presented that vNM utilities can meaningfully be interpersonally compared. In section 5 the conditions are identified under which the aggregation of QALYs over individuals takes the form of "QALY-utilitarianism," i.e. the unweighted summation of QALYs over individuals. Section 6 shows that these conditions inhibit the inclusion of two common types of equity concern: a concern for the fairness of the allocation process, generally referred to as ex ante equity, and a concern for the (final) distributional implications, referred to as ex post equity. Replacing the relevant conditions by alternative conditions allows the

[3]It is important to emphasize that this chapter is concerned with equity concerns over consistently scaled QALYs. Consistent in the sense that QALYs are comparable over individuals. This distinguishes this chapter from for example the chapter by Gafni and Birch (1991) in which the influence of equity considerations on the scaling of the von Neumann Morgenstern utility function is shown and in which algorithms are developed to ensure consistent scaling. However, the two approaches are not completely independent. We will briefly return to this issue in section 6.

[4]These two interpretations are not necessarily mutually exclusive.

inclusion of ex ante and ex post equity concerns. In sections 7 and 8 three procedures are proposed by means of which equity concerns can be captured in cost utility analysis. The procedures described in section 7 only address ex post equity. In section 8 a procedure is described that simultaneously takes into account ex ante equity and ex post equity. Section 9 contains concluding remarks. The appendix contains proofs of results presented in this chapter.

9.2 Interpretations of QALYs

The definition of the number of QALYs for an individual, as given by Pliskin, Shepard and Weinstein (1980), is the following:[5]

$$QALY = \sum_{t=1}^{T} u(q_t) \qquad (1)$$

where T stands for the individual's remaining life time and $u(q_t)$ is the quality weight of health state q_t. At least two interpretations have been distinguished in the literature[6] as to what the number of QALYs represents: *QALYs as vNM utilities* and *QALYs as measures of health*. According to Torrance (1986): "In one approach health state utilities are claimed to be utilities obeying the axioms of von Neumann Morgenstern utility theory.....In the other approach...health state utilities are claimed to measure the overall quality of life" [p.27].

With respect to the first interpretation, QALYs as vNM utilities, conditions have to be imposed on the individual preference relation to ensure that a QALY is a valid vNM utility. Criticism that decision making based on QALYs may not accurately reflect individual preferences is based on the presumption that ideally a QALY should be a vNM utility.[7] In the interpretation of QALYs as vNM utilities, we abstract from the discussion whether the conditions that equate QALYs and vNM

---

[5]One may object against this formulation that it is unnecessarily simple and that for example discounting should be allowed for. However, this simple representation does not imply a loss of generality in terms of the results of this chapter: all results carry over straightforwardly if a more general expression is substituted.

[6]Nord (1994) provides a third interpretation: QALYs as a social value. We will not consider this interpretation for the obvious reason that in this interpretation aggregation plays no role.

[7]E.g. Mehrez and Gafni (1989); Loomes and McKenzie (1989).

utilities are reasonable and it is simply assumed that the individual preference relations satisfy these conditions.

The second interpretation, QALYs as measures of health, is rooted in the extra-welfarist tradition which originates from Sen (1979) and which has been applied to health by Culyer (1989). Wagstaff (1991): "Though utility theory is frequently used in the derivation of quality of life scores, it is used simply to measure people's health rather than the utility they derive from it" [p.23]. Part of the appeal of this latter approach stems from the fact that the comparability of "QALYs as utilities" across individuals may be problematic.

It is not our aim to decide which of these two interpretations is most appropriate. The equity algorithms presented in sections 7 and 8 have been developed with the intention to be applicable under both interpretations. However, equity considerations relate to comparisons between individuals and therefore it has to be established first whether QALYs can be aggregated in both interpretations. The common way to aggregate QALYs is by unweighted summation. Because of this, the QALY approach has been criticized as embodying a return to classical, or Benthamite, utilitarianism. Wagstaff (1991) has argued that in the interpretation of QALYs as measures of health this criticism does not stand scrutiny. Classical utilitarianism focuses on the aggregation of utilities whereas the QALYs as measure of health approach mainly sees QALYs as reflecting characteristics of people without being concerned with the utility they derive from these characteristics. The idea behind this line of argument is that characteristics do not face problems of measurability and comparability across individuals. In the sequel of this chapter this view is taken for granted. It is assumed that in the interpretation as a health measure, QALYs can indeed be aggregated across individuals and that the equity algorithms to be developed later can be applied to QALYs as measures of health.

The assertion made in this chapter that QALYs as vNM utilities can also be meaningfully aggregated requires clarification. The question whether vNM utilities are interpersonally comparable and do have a meaning in social welfare analysis has provoked much debate over the past five decades. In the next section we will establish that to incorporate equity considerations into cost utility analysis full interpersonal comparability of utilities is necessary. In section 4 a rationale is given for why QALYs as vNM utilities can be considered to be fully interpersonally comparable.

9.3 Aggregation of utilities

Under classical utilitarianism social welfare was set equal to the sum of intuitively measurable and comparable individual utilities. These individual utilities were simply assumed to exist, no attention being paid to their origin. This concept of utility and social welfare was challenged by the Pareto school, which claimed that utility is an ordinal concept, reflecting only the individual ordering of outcomes and being incomparable across individuals. Arrow's work on social choice lies within this Paretian tradition. In deriving his celebrated impossibility theorem,[8] Arrow defined a social welfare function (SWF) as a functional relation specifying a social ordering for any given n-tuple of individual orderings. By using only ordering information Arrow deliberately limited the informational framework, excluding all information on preference intensities. Arrow showed that if the number of individuals is finite and if the number of social states is greater than two, no SWF can satisfy the following four conditions: *(i) unrestricted domain:* the SWF should work for all logically possible individual orderings; *(ii) weak Pareto:* if every individual strictly prefers allocation x to allocation y then society should strictly prefer x to y; *(iii) non-dictatorship:* there is no individual such that social preference is completely determined by the preferences of this individual regardless of the preferences of all other individuals in society; *(iv) independence of irrelevant alternatives:* social preference between two allocations should be independent of other allocations. The requirement of independence of irrelevant alternatives excludes all information about other allocations and thereby inhibits the use of any information other than the individual orderings over x and y. Using information on cardinal utility depends on the scaling of the utility function and this necessarily involves taking into account other alternatives.

Various attempts have been undertaken to escape from Arrow's impossibility theorem by weakening his conditions. In this chapter we consider the enrichment of the informational base of Arrow's social choice approach. A social welfare functional (SWFL) is defined as a rule that specifies exactly one social ordering for any given n-tuple of real-valued individual utility functions. Let L_i be defined as the set of individual utility functions that are informationally equivalent, i.e. that provide the same information on individual preferences. For example, given Arrow's assumptions, all individual utility functions that are positive monotonic transformations are informationally equivalent. If individual utility is cardinally measurable, then elements of the

[8] See Arrow (1950,1951a,1963). In Arrow (1950, 1951a) the domain restriction was not defined tight enough as was pointed out by Blau (1957). The version in Arrow (1963) is the best known version.

set L_i of informationally equivalent utility functions are positive linear transformations of each other: $U_i = a + bU_i'$, $a \in \mathbb{R}$; $b > 0$; U_i, $U_i' \in L_i$.

A measurability set L is defined as the set of all possible combinations of the n-tuples of informationally equivalent individual utility functions. Depending on the assumptions about interpersonal comparability, the measurability set can be restricted. Combining the measurability assumptions about individual utilities with the assumptions about interpersonal comparability defines the measurability-comparability set L^*. In Arrow's framework of social choice, where no interpersonal comparability is assumed and only the information revealed by individual orderings is incorporated, L^* consists of all individual utility functions that are positive monotonic transformations of each other. Sen (1970b, 1977b) distinguishes several other measurability-comparability combinations:

- *cardinal non-comparability*: L^* consists of all individual utility functions that are unique up to positive linear transformations.
- *ordinal level comparability*: L^* consists of all individual utility functions that are unique up to *similar* positive monotonic transformations.
- *cardinal unit comparability*: L^* consists of all individual utility functions that are unique up to location and common scale: $U_i = a_i + bU^*_i$, $a_i \in \mathbb{R}$ (the set of real numbers); $b > 0$.
- *cardinal full comparability*: L^* consists of all individual utility functions that are unique up to common location and common scale: $U_i = a + bU^*_i$, $a \in \mathbb{R}$; $b > 0$.

Lemma 8*2 in Sen (1970) shows that assuming cardinal non-comparability is not sufficient to solve Arrow's impossibility result. However, the other three informational frameworks are sufficient to remove the dilemma posed by Arrow's theorem. Clearly, it is interpersonal comparability that is crucial in enriching the informational basis of social choice.

In cost utility analysis the calculation of the net advantage of one program over another is of interest. For such an analysis to be relevant, units should be comparable. Location need not necessarily be common to all individuals, since in calculating net advantages the individual-specific locations are subtracted away and play no role in determining the relative effectiveness of programs. Suppose for example that for a particular n-tuple of individual utility functions program x is preferred to program y. That is, $\Sigma[U_i(x_i) - U_i(y_i)] > 0$. But also $\Sigma[a_i + bU_i(x_i) - a_i + bU_i(y_i)] = b\Sigma[U_i(x_i) - U_i(y_i)] > 0$, $a_i \in \mathbb{R}$ and $b > 0$, and thus adding individual specific constants does not influence the relative effectiveness of programs. If b would be individual specific,

which corresponds with cardinal non-comparability, x might no longer be preferred to y. This suggests that in cost utility analyses we need only impose cardinal utility functions that have their scale in common, i.e. cardinal unit comparability. It seems not necessary to assume level comparability. However, several authors[9] have shown that if cardinal unit comparability is assumed rather than cardinal full comparability, slightly strengthened versions of Arrow's conditions imply that the only possible SWFL is the utilitarian one. In such an informational framework, simply aggregating the number of QALYs over the relevant population is unobjectionable. This result was to be expected. The notion of equity involves special consideration being given to the badly-off and this necessarily involves bringing in comparisons of utility levels. Given that the starting point of this chapter was a concern for the equity consequences of QALY-based decision making, a framework has to be imposed that allows such concerns to be justified. That is, a rationale must be given for assuming cardinal full comparability.

9.4 von Neumann Morgenstern utilities

In case the vNM axioms hold individual utility functions are cardinal, i.e. unique up to positive affine transformations. One may suggest therefore to use individual vNM utilities as an input in the social welfare functional. Taking individual vNM utilities as the basis from which social welfare judgements are to be derived has first been proposed by Harsanyi (1955). Harsanyi's position has been severely contended though. The essence of the criticism being that vNM utilities are inextricably bound to situations involving risk. Arrow (1951a, p.10): "....it [vNM utility theory] has nothing to do with welfare considerations, particularly if we are interested primarily in making a social choice among alternative policies in which no random elements enter. To say otherwise would be to assert that the distribution of the social income is to be governed by the tastes of individuals for gambling."

One can respond to such criticism in one of two ways. The first type of answer acknowledges that vNM utilities are only relevant in the context of risk, but asserts that health decision making typically involves risk and that, therefore, vNM utilities do have relevance in this context [e.g. Ben-Zion and Gafni (1983)]. The second type of response challenges the assertion that vNM utilities only have relevance in the context of risk. According to this line of reasoning, cardinal utility has a meaning independent of risk. That cardinal utility has a meaning independent of risk has been criticized by

[9]E.g. d'Aspremont and Gevers (1977), Sen (1977b), Deschamps and Gevers (1978).

Arrow (1951b) who writes about cardinal utility under certainty: "...which is a meaningless concept anyway [p.425]." Similar views have been expressed by Savage (1954), Ellsberg (1954), Luce and Raiffa (1957) and Fishburn (1989). Harsanyi (1987) on the other hand asserts that: "In fact, people's vNM utility functions are an important piece of information for welfare economics and ethics because they are natural measures for the intensity of people's desires, preferences and wants (p.546-547)." In a recent chapter, Wakker (1994) provides a defense for a unified notion of utility that does not need risk for its existence, but that has relevance for risk. Wakker observes that the development of expected utility theory by von Neumann and Morgenstern was motivated by their desire to obtain a cardinal utility that is relevant to game theory. The same cardinal utility, the expectation of which represents individual choices over lotteries over outcomes, is used as a unit of exchange between players in a game. Wakker (p.8): "I think the applicability of risky utility as means of exchange between players is as questionable as its applicability to welfare theory, or any other case of decisions under certainty." It cannot be excluded that vNM had in mind one notion of utility for the entire economic science. This viewpoint, that vNM utilities do indeed have relevance in other contexts than risk, underlies the discussion of QALYs as vNM utilities in this chapter.

Having provided a rationale for using cardinal (vNM) utilities as the foundation of social welfare judgements, the question remains how interpersonal comparability can be ensured given that individual utilities are unique only up to positive linear transformations and given that scaling up the utility of one individual, while keeping the utilities of the other individuals constant, may alter the outcome of the social choice problem. Hildreth (1953) has suggested to consider two specially defined outcomes X and Y, such that everyone prefers X to Y, and to assign predefined real values to these social states. This makes individual utility functions interpersonally comparable. In fact this approach is typically used in cost utility analysis. The general approach to aggregation, as outlined for example by Williams (1981) and Torrance (1986), is to assign a utility of zero to death and a utility of one to normal or full health and to regard a year of healthy life as being of equal intrinsic value to everyone. Gafni and Birch (1991) have argued that it may be more in line with existing practice to assume that a life in full health has equal value for everyone. In their approach the vNM utility function is scaled such that a life in full health receives utility one and immediate death utility zero. This approach guarantees that individual utilities are consistently scaled and are interpersonally comparable.

Summarizing, the above discussion establishes a rationale for aggregating QALYs as (vNM) utilities. Section 3 showed the need for cardinal fully comparable utilities if we are to allow distributional considerations to play a role in cost utility

analysis. vNM utility theory establishes cardinality of the individual utilities. Following Wakker's argument a case can be made for the assertion that vNM utilities do indeed have relevance in the context of welfare judgments. Finally, by Hildreth's approach, which is typically followed in cost utility analysis, a consistent scaling procedure emerges, which ensures that individual vNM utilities can be interpersonally compared in a meaningful way.

9.5 QALY utilitarianism

9.5.1 Notation and structural assumptions

This subsection introduces notation and structural assumptions. Denote the set of QALY allocations by X. A typical element of the set X is a vector $x=(x_1,\ldots,x_n)$ representing an allocation of QALYs resulting from the implementation of a health care program with each x_i indicating the number of QALYs received by individual i and n being the number of individuals affected by the program. Assume without loss of generality that for each individual the possible number of QALYs is non-negative, i.e. $X\in\mathbb{R}_+$. We are interested in the social preference relation $\succsim$ over the set of QALY allocations, meaning "at least as good as". Let $\succ$ and $\sim$ denote its asymmetric and symmetric part respectively. Throughout $\succsim$ is assumed to be a weak order. That is, $\succsim$ is complete, either $x\succsim y$ or $y\succsim x$ or both, and transitive, if $x\succsim y \& y\succsim z$ then $x\succsim z$. Moreover, $\succsim$ is continuous. Continuity of the preference relation guarantees that if a real-valued function is defined over X, this function has an interval as its image.

Denote by $x_{-i}v_i$ the vector x with coordinate i (the number of QALYs individual i receives) replaced by v_i: $x_{-i}v_i=(x_1,x_2,..x_{i-1},v_i,x_{i+1},\ldots,x_n)$. Let A be a subset of the individuals affected by a health care program: $A\subset I=\{1,2,\ldots,n\}$. Then $x_{-A}v_A$ denotes the vector x in which for all individuals in subset A x_i is replaced by v_i. For example if $A=\{1,2,3\}$, then $x_{-A}v_A=(v_1,v_2,v_3,x_4,\ldots,x_n)$. Denote by $\succsim_i$ the individual preference relation "at least as good as". As before, $\succ_i$ and $\sim_i$ are defined as the asymmetric and symmetric part of $\succsim_i$ respectively.

Let Z be a set of probability distributions over the set of QALY allocations X. A typical element of Z is $(p^1,x^1;\ldots;p^m,x^m)$ where allocation x^j occurs with probability p^j and m can be any natural number. Let $\succsim_z$ be a social preference relation defined on Z. Throughout the chapter it is assumed that individual preference relations over probability distributions satisfy the von Neumann Morgenstern (vNM) axioms. In the formulation by Jensen (1967), a preference relation $\succsim'$ satisfies the vNM axioms if:

(i) $\succsim'$ is a weak order; *(ii)* vNM independence: $P \succsim' Q \Leftrightarrow (\mu P + (1-\mu)R) \succsim (\mu Q + (1-\mu)R), 0 < \mu < 1$ *and* $P,Q,R \in Z$; *(iii)* Jensen continuity: $(P \succ' Q, Q \succ' R) \Rightarrow \kappa P + (1-\kappa)R \succ' Q$ *and* $Q \succ' \rho P + (1-\rho)R$ for some $\kappa,\rho \in (0,1)$. If a preference relation satisfies the vNM axioms, a cardinal real-valued utility function exists the expected value of which represents the preference relation.

9.5.2 Derivation of QALY utilitarianism

We will derive (QALY) utilitarianism by adding a condition to the axiomatic framework of Harsanyi (1955) in which a partial characterization of utilitarianism is given. The method of proof differs from the one given by Harsanyi in that use is made of a result developed by Fishburn (1965). Further by using a theorem from Maas and Wakker (1994) the utility function is shown to be continuous. Continuity is important to establish. If QALYs (real numbers) are added up across individuals the social utility function is implicitly assumed to be continuous. However Harsanyi's result does not imply this.

Harsanyi not only requires *individual* preferences to satisfy the vNM axioms, as has been assumed in subsection 5.1, but also required *social* preferences to satisfy the vNM axioms. According to Harsanyi the vNM axioms are essential requirements of rationality, much in the same spirit as Arrow considered weak ordering to be a basic requirement of rationality of social preferences. Further Harsanyi imposed the following condition:

Condition H: If two alternatives, defined by probability distributions over the set of outcomes, are indifferent from the standpoint of every individual, then they are also indifferent from a social standpoint.

As shown by Harsanyi (theorem V), these three conditions allow the derivation of the SWFL as a weighted sum of the individual utilities:

$$U(x) = \sum_{i=1}^{n} \lambda_i U_i(x_i) \qquad (2)$$

This is not a full characterization of QALY-utilitarianism, given that scaling factors λ_i may differ between individuals and utility functions are individual-specific. QALYs are assumed to be similar across individuals. Therefore a condition has to be added to ensure this similarity.

A permutation π of the n individuals is a function specifying a rearrangement of the individuals. Denote by $\pi(i)$ the permuted value of i. Now consider the following condition:

Condition A (anonymity): $(U_i) \sim_z (U_{\pi(i)})$ *for all* $(U_i) = (U_1, \ldots, U_n)$ *and permutation functions* π *on* $I = \{1, \ldots, n\}$.

Condition A ensures that social preference is independent of who gets which utility/QALY. By condition A, if there is one additional QALY to be divided between two individuals with a similar endowment of QALYs, then society should have no preference as to which individual will receive this additional QALY. However, condition A is weaker than what is referred to in the cost utility literature as "a QALY is a QALY no matter who gets it". According to the latter, society should in any situation be indifferent with respect to who gets a QALY. Condition A only says that in case one QALY allocation is a permutation of another, indifference should hold. For example, suppose a program has resulted in a QALY allocation $(3,1)$, i.e. individual a has received three QALYs and individual b has received one QALY, and one more QALY is to be allocated. Then by the argument that "a QALY is a QALY no matter who gets it" society should be indifferent between allocations $(4,1)$ and $(3,2)$. However, condition A does not provide guidance with respect to social preference between $(4,1)$ and $(3,2)$. Condition A asserts that if society prefers $(3,2)$ to $(4,1)$ then it should also prefer $(2,3)$ to $(1,4)$ when the initial allocation is $(1,3)$: by condition A $(2,3) \sim_z (3,2)$; we know that $(3,2) \succ_z (4,1)$; applying condition A once again gives $(4,1) \sim_z (1,4)$ and thus by transitivity $(2,3) \succ_z (1,4)$. Imposing condition A on top of Harsanyi's conditions is necessary and sufficient for QALY utilitarianism.

Theorem 9.1: The following two statements are equivalent:
(i) The social preference relation $\succeq_z$ *can be represented by QALY utilitarianism:*

$$U(x) = \sum_{i=1}^{n} U(x_i) \qquad (3)$$

(ii) both individual and social preferences satisfy the vNM axioms and moreover conditions H and A hold.

Furthermore U is continuous and unique up to positive linear transformations.

A proof of this result can be found in the appendix.

9.6 Ex ante versus ex post equity

Theorem 9.1 has been derived by imposing four conditions: that individual preferences satisfy the vNM axioms, that social preferences satisfy the vNM axioms, condition *H* and condition *A*. In the remainder of the chapter we continue to require that individual preferences satisfy the vNM axioms. Particularly during the last two decades much empirical evidence has been presented that descriptively individual preferences frequently violate these axioms. Normatively the axioms still have considerable force and are appealing enough to adhere to. We will also continue to assume that condition *A* holds. Condition *A* asserts that the identity of a QALY recipient should play no role in health decision maker, and this appears a reasonable condition to impose. Condition *A* ensures that the principle that a life in full health should be equal for all individuals holds, and thereby allows consistent scaling of the utility functions according to the equity principles developed by Gafni and Birch (1991). Moreover, as the example in the previous section shows, condition *A* does not predict choice with respect to every allocation and therefore allows additional equity principles to be imposed.

We will examine the consequences of relaxing the two remaining conditions, that social preferences satisfy the vNM axioms and condition *H*. The restrictiveness of these assumptions can be illustrated by means of an example. Consider two individuals (or equivalently two groups of individuals) and two possible states of the world, *X* and *Y*, each with a probability of occurrence of *0.5*. This probability is known to both individuals. Consider the following three health care programs each resulting in different QALY allocations:

Program	State X	State Y	Expected Utility
1	(1,0)	(1,0)	1
2	(1,0)	(0,1)	1
3	(1,1)	(0,0)	1

Under the assumptions being made, by theorem 9.1, the decision maker should be indifferent between the three health care programs, given that the expected utilities of the three programs are equal. However, it is conceivable that the decision maker will prefer programs 2 and 3 to program 1 given that the former two programs offer both

individuals a possibility of receiving a QALY, whereas program 1 denies the second individual the possibility of receiving a QALY. Diamond (1967)[10] has argued that it is essentially vNM independence that requires indifference to hold in the above example. This is most easily seen by comparing programs 1 and 2. Under condition A, the decision maker is indifferent between the outcomes of the two programs when state Y occurs. However, under state X the outcomes of the two programs are equal and therefore, by vNM independence, overall indifference should prevail. On the other hand, if the decision maker is concerned with the fairness of the allocation process, generally referred to as ex ante equity, program 2 should be chosen, because this gives both individuals a possibility of obtaining a QALY. Incorporating ex ante equity concerns means dismissing the requirement that social preferences satisfy the vNM axioms. Incorporating ex ante equity considerations can be ensured by imposing the following ex ante equity condition on the social preference relation:

Condition E: If $p_k=q_k$ *for all* $k\in I\setminus\{i,j\}$; $p_i+p_j=q_i+q_j$ *and* $|p_i - p_j| < |q_i - q_j|$ *then* $P\succ_z Q$.[11]

where P and Q are lotteries over X and the p_i's and q_i's are marginal probabilities, indicating the probability that individual i receives a given amount of QALYs, Q^c. In words condition E says the following. Suppose all individuals, other than i and j, have the same marginal probability of receiving Q^c under two health care programs (in the above example Q^c is equal to one). Taken together i and j have the same marginal probability of receiving Q^c, but in one program this marginal probability is more equally divided between the two individuals (in the situation described by condition E this is the program giving rise to probability distribution P). Then, by condition E, the program with the more equal distribution of marginal probabilities over i and j is to be preferred to the one that leads to a less equal distribution of marginal probabilities over i and j.

Condition E is not incompatible with condition H. Condition E dictates how differences in marginal probabilities should affect social preference, whereas condition H dictates how equality of marginal probabilities should affect social preference.

It is also conceivable that, in choosing between programs 2 and 3, a social decision maker prefers program 3 over program 2, given that program 3 guarantees an equal distribution of QALYs under both states of the world, whereas program 2

---

[10]See also Sen (1976), Broome (1982), Ulph (1982). For a counterargument see Harsanyi (1975).

[11]Condition E is comparable to Fishburn's (1984) axiom of risk-sharing equity in the context of public risk evaluation. See also Fishburn and Straffin (1989).

necessarily leads to a situation of inequality. This preference is determined by a concern for the final distribution of QALYs, often referred to as ex post equity. Let $p_{i\text{-}j}$ denote the probability of the event that only individual i gets a QALY and let p_{i+j} denote the probability of the event that both individual i and individual j receive a QALY. In the example above, $p_{i\text{-}j}$ is *0.5* for program 2 and 0 for program 3, whereas p_{i+j} is *0* for program 2 and *0.5* for program 3. Incorporating a concern for ex post equity can be established by imposing the following condition on the social preference relation:

Condition P: *If* $P = Q$, $P, Q \in Z$ *apart from* $p_{i\text{-}j} = q_{i\text{-}j} \cdot \gamma$, $p_{j\text{-}i} = q_{j\text{-}i} \cdot \gamma$, $p_{i+j} = q_{i+j} + \gamma$, *with* $\gamma > 0$, *then* $P \succ_z Q$.[12]

In the situation described by condition *P*, $p_i = q_i$, so by condition *H*, $P \sim_z Q$. However, probability distribution *P* offers a greater probability of individuals i and j both receiving a QALY. Therefore, by condition *P*, $P \succ_z Q$. Thus, incorporating ex post equity considerations in health care decision making means rejection of condition *H*. Condition *H*, innocuous as it may appear, has the effect of making social choice dependent on individual preferences only. The condition leaves no room for supra-individual interests. Incorporating distributional concerns therefore means relaxing the condition that social choice depends only on individual preferences and allowing complementarity between individual utility/health levels. In the next section we discuss two approaches to incorporate such complementarity.

9.7 Ex post equity algorithms for QALY aggregation

9.7.1 A multiplicative social utility function

As has been observed in the proof of theorem 9.1, condition *H* is equivalent to additive independence (Fishburn, 1965, 1970). Therefore, a way to introduce complementarity seems to translate generalizations of additive independence, known from the literature on multi-attribute utility theory, to the context of social choice. One possibility is to impose the analogue of mutual utility independence on the social

[12]This condition is the converse of Keeney's (1980) assumption of catastrophe avoidance. Fishburn (1984) and Fishburn and Straffin (1989) have developed similar "common fate equity" axioms for the context of public risk evaluation

preference relation.[13] Mutual utility independence is a preference condition that is entirely formulated in terms of lotteries on outcomes. We deviate slightly from this approach by making as little use of lotteries as possible in the conditions imposed on the social preference relation. The main motivation underlying our approach is that lotteries are highly artificial constructs that are typically not available in real world health decision situations. An additional motivation to use the independence condition *SE* as stated below is that this condition is more common in social choice theory and that a justification for imposing it has been given.[14]

Condition SE: The social preference relation $\succsim$ *satisfies condition SE if for all QALY allocations* $x, x', y, y' \in X$, *for all subsets of individuals* $A \subset I=\{1,\ldots,n\}$:

$$[v_{.A}\, x_A \succsim v_{.A}\, y_A] \Leftrightarrow [w_{.A} x_A \succsim w_{.A} y_A]$$

By condition *SE*, individuals who are indifferent between two QALY allocations, the individuals who are not in subset A, exert no influence on social preference. Condition *SE* is the analogue of the "sure thing principle" in the context of decision making under uncertainty [Savage, 1954] and of "complete strict separability" in consumer theory [Blackorby et al., 1978]. Condition *SE* underlies the current practice of using incremental analysis in cost utility analysis [cf. e.g. Drummond et al., 1987]. Incremental analysis prescribes to calculate the net advantage of one program over another. The implication of this is that if two programs produce the same amount of QALYs for certain individuals, then these individuals do not influence the outcome of the analysis. This is exactly what condition *SE* asserts.

Condition *SE* is formulated under certainty. However, since we will present a representation for the social preference relation under risk, we have to impose a condition which is defined with respect to preferences under risk. Consider the following condition:

Condition UI: Let B be a subset of individuals, i.e. $B \subset I=\{1,\ldots,n\}$, *let y be a particular constant QALY allocation,* $y \in X$, *and let* $\succsim_{z|y}$ *be the preference relation defined over probability distributions on* IR_+^B *by fixing the values of those individuals outside subset B (I-B) at levels identical to those of y. B is utility independent if* $\succsim_{z|y}$ *is independent of the constant value at which y is fixed.*

---

[13]For a definition of mutual utility independence see for example Keeney and Raiffa (1976, p.289).

[14]For a defense see Fleming (1952); Deschamps and Gevers (1978); and Sen (1976, 1977b)

In the special case where all probability distributions are degenerate, i.e. one outcome results with probability one, condition *UI* is equivalent to condition *SE*. If condition *UI* holds for all subsets of individuals *B*, mutual utility independence holds. However, in combination with condition *SE* it is not necessary to impose condition *UI* for all subsets of individuals. It is sufficient to impose that *UI* holds for *one* individual. Thus, if all other *n-1* individuals are indifferent between two QALY allocations, then social preferences for lotteries on these two allocations are governed by the preferences of this particular individual. Denote this condition as UI^1. Condition UI^1 only holds when all other individuals are indifferent. The relevant individual can therefore not be considered to be a dictator in Arrow's sense. Condition UI^1 is a somewhat artificial condition. However, it is not very restrictive in terms of the social preference relation. If the one individual for who condition UI^1 holds, is the individual who is worst off in terms of health, then it seems defendable to impose that, in case all other individuals are indifferent, social preferences under risk should be governed by the preferences under risk of this individual.

Theorem 9.2: The following two statements are equivalent:
(i) the social preference relation can be represented by:

$$U(x) = (1/\lambda)\prod_{i=1}^{n} [\lambda U(x_i) + 1] - (1/\lambda) \qquad (4)$$

where $U(x)$ is a continuous social utility function, unique up to positive linear transformations and scaled between 0 and 1, the $U(x_i)$ are identical additive utility functions, that can be interpreted as (re-scaled) QALYs and λ is a scaling constant, that is not equal to zero.
(ii) both individual and social preferences satisfy the vNM axioms; social preferences satisfy conditions A, SE and UI^1. If condition P also holds then $\lambda > 0$.

A proof can be found in the appendix.

The scaling parameter λ reflects the influence of complementarity. For example, for two individuals equation (4) reduces to:

$$U(x) = U(x_1) + U(x_2) + \lambda U(x_1)U(x_2) \qquad (5)$$

If $\lambda > 0$, complementarity increases social utility, which is the effect of imposing condition *P*.

Consider the example for the two individuals[15] and the three health care programs described in section 6. Recalculating the social utility of health care programs 1, 2 and 3 gives 1, 1 and $1+0.5\lambda$ respectively. Thus, under condition *P*, program 3 is now preferred, which is consistent with the (imposed) preference for ex post equity. Indifference still holds for programs 1 and 2, since they have the same distributional implications. Indifference between programs 1 and 2 reflects the fact that ex ante equity has not been taken into account.

The value of λ reflects policy views on equity. These policy views can partly be expressed by conditions such as condition *P*, but to determine the relative weight given to aggregating the individual QALYs (the efficiency side) and to complementarity between individual QALYs (the ex post equity side) requires the policy maker to make explicit his choices with respect to the equity-efficiency trade-off. Trading off attributes is common practice in multi-attribute utility theory and the tools of multi-attribute utility theory can be of great help in eliciting preferences between efficiency and equity in health care.

Under condition *A* one trade-off question is sufficient to determine λ. As an illustration, consider again the example of two individuals. In theorem 9.2, the $U(x_i)$ are re-scaled QALYs (for more details see the proof of theorem 9.2 in the appendix): $\lambda_i U'(x_i)$ where all λ_i are equal and positive and $U'(x_i)$ indicates the number of QALYs each individual receives. For the purpose of the theorem this was no problem given that vNM utility functions are unique up to positive linear transformations. However, to calculate λ we need to determine λ_i. This can be done by asking the policy maker to give an indifference probability for the choice between $(1,0)$[16] with certainty and a gamble with outcomes $(1,1)$ with probability p and $(0,0)$ with probability $(1\text{-}p)$. Suppose the policy maker's indifference probability is 0.4. Scale $U(x)$ such that $U(1,1) = 1$. Then, substituting values in equation 5, $1 = 0.4*1 + 0.4*1 + \lambda*0.4*1*0.4*1$. This gives $\lambda = 1.25$. It is conceivable that a policy maker cannot in every situation specify exact values for λ, but only a range of values. In that case it seems sensible to include this range of values for λ in sensitivity analyses.

Finally, the multiplicative social utility function, as derived above, only incorporates equity concerns to a limited extent. By condition *SE*, indifferent individuals do not exert an influence on social preference. In a situation where the non-indifferent

---

[15]To be formally correct, for two individuals a stronger condition than *SE* has to be imposed: the analogue of Wakker's (1989) hexagon condition which will be discussed in section 8.

[16]Or $(0,1)$ which is under condition A equivalent to $(1,0)$.

individuals are already in a good health state but the indifferent individuals are in an appalling health state, a policy maker may prefer the non-indifferent individuals not to receive more QALYs in order to prevent a more unequal distribution of health (utility). Such equity concerns cannot be accommodated by the proposed multiplicative social utility function. In the next section we propose a social utility function which is able to embrace ex post equity concerns in a more comprehensive way.

9.7.2 A two component social utility function

Continue to assume that social preferences satisfy the vNM axioms. Therefore, as in section 5 and in subsection 7.1, the social preference relation is defined over probability distributions. We propose a method that allows the decision maker to simultaneously consider the maximization of QALYs, that can both be interpreted as health and utility, and the distribution of these QALYs, that is ex post equity. The idea is to assess a two component social utility function $U(y) = U(y_1, y_2)$, the components of which are the total number of QALYs (y_1) and a real valued summary index reflecting the ex post distribution of these (y_2). The set of outcomes, Y, is assumed to satisfy certain structural assumptions.[17] The assessment of such a two component multi-attribute utility function is greatly facilitated if the following assumption can be accepted:

Condition TCI (two component independence): If two lotteries induce the same probability distribution over Y_1 (total number of QALYs gained) and the same probability distribution over Y_2 (the summary index reflecting the ex post distribution), then these lotteries are indifferent.

This condition is similar to additive independence (and to condition H for the case of two individuals) and guarantees, in combination with the assumption that the social preference relation satisfies the vNM axioms, by theorem 2 in Fishburn (1965) that $U(y)$ is additive:

$$U(y) = \lambda_1 U_1(y_1) + \lambda_2 U_2(y_2) \qquad (6)$$

[17] Y is assumed to be a Cartesian product of Y_1 and Y_2, $Y_1 \in \mathbb{R}_+ \setminus \{0\}$ and $Y_2 \in \mathbb{R}_+$. The reason that 0 is excluded from Y_1 is that otherwise not every value from Y_1 can be combined with every value from Y_2 and, by consequence, Y cannot be a Cartesian product.

where U, U_1 and U_2 are scaled vNM utility functions, and λ_1 and λ_2 are scaling constants that reflect policy views on the trade-off between "efficiency" in the sense of the maximization of QALYs and ex post equity.

The summary index defined over the ex post distribution should satisfy certain properties. For example, it should be sensitive to a transfer from an individual who is relatively well off in terms of the number of QALYs received from the implementation of a health care program, to an individual who receives less QALYs from this program. An example of such a summary index is Theil's entropy measure [cf. Sen, 1973]:

$$y_2 = \sum_{i=1}^{n} x_i ln(nx_i) \qquad (7)$$

where x_i denotes the share of the total amount of QALYs received by individual i. y_2 increases with inequality in the QALY distribution, therefore, under condition P, $\lambda_2 U(y_2)$ must have a negative sign.

Assume that condition *TCI* holds. Assume further that the utility function for the amount of QALYs is linear and that the utility function for the ex post distribution is equal to Theil's entropy index. Then for the example in section 6 we obtain: $U(program\ 1) = U(program\ 2) = \lambda_1 + \lambda_2 ln2$; $U(program\ 3) = \lambda_1$. Under condition P $\lambda_2 < 0$, and thus program 3 is preferred, consistent with a preference for ex post equity.

Only one trade-off question has to be asked to determine the scaling constants, λ_1 and λ_2. Suppose that a program yields benefits for two groups of individuals and that the maximum amount of QALYs the program can generate is 100. Then the best possible outcome for the policy maker is $(50,50)$: the number of QALYs is maximized and there is no inequality. Scale $U(.)$ such that $U(50,50) = 1$. The worst outcome is $(x,0)$ in which x is infinitesimally small: the number of QALYs is minimized and there is complete inequality. Let $U(x,0)$ be zero. Now λ_1 can be determined by eliciting the policy maker's indifference probability in a choice between (100,0), i.e. the number of QALYs is at its maximum, but inequality is complete, for certain and a gamble giving (50,50) with probability p and (x,0) with probability $(1-p)$. Suppose the policy maker indicates that $p=0.85$. Substituting in equation (6) gives: $U(100,0) = \lambda_1 * 1 = p$. Thus $\lambda_1 = 0.85$ and λ_2 is by consequence equal to 0.15.

If condition *TCI* does not hold, complementarity between y_1 and y_2 has to be introduced in the model. For example, if the social preference relation does not satisfy

condition *TCI*, but does satisfy a somewhat stronger condition than *SE*[18], and does satisfy *UI*[1], then the term $\lambda\lambda_1\lambda_2 U_1(y_1)U_2(y_2)$ should be added to the additive form, reflecting complementarity. In this case one additional trade-off question has to be asked to determine the scaling constants.

9.8 Algorithms incorporating both ex post and ex ante equity

In section 6 it was argued that if a concern for ex ante equity is to be incorporated in social preference, the vNM utility function can no longer be used. Therefore, in this section, rather than taking a preference relation over probability distributions as primitive, we will seek a representation for a preference relation under certainty. In this section we will consider a three component social value function $V(y) = V(y_1, y_2, y_3)$, where y_1, y_2 and y_3 denote the number of QALYs gained, which can be both utilities and health, and real valued summary indices reflecting the ex post equity and the ex ante equity of the QALY allocation process respectively. More specifically we will derive a representation for the value function $V(y_1^{ce}, y_3)$[19] in which y_1^{ce} denotes the certainty equivalent amount of QALYs gained, with the ex post equity held fixed, for probability distributions over y_1 and y_2. For example, if the ex post equity index is fixed at its optimal value corresponding with no inequality, then for every lottery the equivalent number of equally distributed QALYs is determined. Under the assumption that social preferences increase monotonically with the number of QALYs, which seems reasonable and is typically assumed in cost utility analysis, the equivalent number of QALYs will consistently rank order lotteries, a higher number corresponding to more preferred. $V(y_1^{ce}, y_3)$ is equivalent to $V(U(y_1, y_2), y_3)$ in which U is a vNM utility function defined over y_1 and y_2. By means of the social value function the certainty equivalent number of QALYs of a gamble can be traded off against its ex ante equity implications. Thus we continue to assume that social preferences with respect to lotteries over y_1 and y_2 while holding y_3 fixed satisfy the vNM axioms. vNM utility functions are still used to evaluate attributes y_1 and y_2, because ex ante equity is defined in the context of risk and therefore utility functions that are applicable in the context of decision making under risk are called for. Social policy making is essentially a normative decision problem,

---

[18] The hexagon condition.

[19] The set Y is again assumed to be a Cartesian product set. It is assumed that $Y_1 \in \mathbb{R}_+ \setminus \{0\}; Y_2, Y_3 \in \mathbb{R}_+$. In combination with the assumption that $\succeq$ on Y is a weak order, this guarantees the existence of $V(Y)$. If $\succeq$ on Y is moreover assumed to be continuous, then $V(Y)$ will be continuous.

and to date there is no theory that challenges expected utility theory as a normative theory of decision making under risk. Diamond's objection against vNM utility theory concerned its implications for ex ante equity. Diamond's argument does not conflict with the use of vNM utility functions to evaluate y_1 and y_2. We will describe the preference conditions that make it possible to represent $V(Y)$ by the following simple expression:

$$V(y) = \kappa_1 V_1[\lambda_1 U_1(y_1) + \lambda_2 U_2(y_2)] + \kappa_3 V_3(y_3) \qquad (8)$$

where U_1 and U_2 are (scaled) vNM utility functions, and V_1 and V_3 are (scaled) value functions.

Consider the following preference condition:

Hexagon condition: if $[(y_1^{ce'}, y_3) \sim (y_1^{ce}\ y_3\ ')\ \&\ (y_1^{ce''},\ y_3) \sim (y_1^{ce}\ ', y_3')\ \&\ (\ y_1^{ce}\ ',\ y_3'\ ) \sim (y_1^{ce},\ y_3'')]$ *then*
$(y_1^{ce''}, y_3') \sim (y_1^{ce'}, y_3'')$

Suppose that $y_1^{ce}\ '' \succ y_1^{ce}\ ' \succ y_1^{ce}$ and that $y_3'' \succ y_3' \succ y_3$. By the first two indifferences in the hexagon condition, both the utility difference between $y_1^{ce'}$ and y_1^{ce} and the utility difference between $y_1^{ce''}$ and $y_1^{ce'}$ are just sufficient to compensate the utility difference between y_3' and y_3. Then the third and the (implied) fourth indifference assert that if the utility difference between between $y_1^{ce'}$ and y_1^{ce} is also just sufficient to compensate the utility difference between y_3'' and y_3', then the utility difference between $y_1^{ce''}$ and $y_1^{ce'}$ should also be just sufficient to compensate the utility difference between y_3'' and y_3'. The hexagon condition is, under transitivity of the indifference relation, implied by the Thomsen condition, which has been more commonly used as a characterizing condition for an additive two attribute utility function [e.g. Debreu, 1960].

Given that the hexagon condition holds, a preference relation $\succsim_A$ can be defined over probability distributions on y_1 and y_2, while fixing y_3 at some constant reference value. This preference relation is assumed to satisfy the vNM axioms and condition *TCI* (two component independence).

Then the following result can be stated:

Theorem 9.3: The following are equivalent:
(i) V(Y) can be represented by equation (8)
(ii) the social preference relation on Y is a continuous weak order that satisfies the hexagon condition, and $\succsim_A$ *satisfies the vNM axioms and condition TCI.*

Furthermore, V, U_1, U_2, V_3 *are continuous and unique up to positive linear transformations. The* λ_i*'s are scaling constants.*

A proof of this theorem can be found in the appendix.

The summary index y_3, reflecting ex ante equity, should be sensitive to changes between individuals in the marginal probability of obtaining a given amount of QALYs, Q^c. The following index possesses this property:

$$y_3 = (1/n)\sum_{i=1}^{n} (q_i - q_m)^2 \qquad (9)$$

where q_i denotes the marginal probability of individual i receiving Q^c and q_m denotes the mean probability of receiving Q^c. A more equal distribution of marginal probabilities leads to a lower value for the summary index. Therefore, under condition E, $\kappa_3 V_3(y_3)$ should be negative. The amount of QALYs with respect to which q_i and q_m are defined should be chosen according to what the policy maker believes individuals are entitled to. For example, it may be the policy maker's conviction that every individual should have an equal probability of receiving a life in full health. In that case, q_i denotes the individual probability of obtaining a life in full health.

Suppose with respect to the example of section 6 that the conditions of theorem 9.3 hold and that U_1, U_2, V_1 and V_3 are identity functions, i.e. $U_i(y_i) = y_i$, with y_2 and y_3 as in equations (7) and (9) respectively. Then $V(program\ 1) = \kappa_1(\lambda_1 + \lambda_2 ln2) + 0.25\ \lambda_3$; $V(program\ 2) = \kappa_1(\lambda_1 + \lambda_2 ln2)$; $V(program\ 3) = \kappa_1\lambda_1$. Imposing conditions E and P has the effect of making λ_2 and λ_3 both negative. Therefore, under conditions E and P the resulting ranking of the health care programs is: $3 \succ 2 \succ 1$.

The assessment of the scaling constants follows from a procedure similar to the one outlined at the end of section 7.2. The only difference is that in this case two trade-off questions have to be asked given that there is one additional scaling constant (one of the κ's) to assess. The additive value functions V_1 and V_3 can be assessed along the lines sketched for example in section 3.7 in Keeney and Raiffa (1976).

Condition *TCI* is a restrictive condition. If a policy makers wants to introduce complementarity between $U(y_1)$ and $U(y_2)$, condition *TCI* can for example be

replaced by the weaker condition of mutual utility independence between y_1 and y_2. If condition *TCI* is replaced by mutual utility independence, but the other conditions of theorem 9.3 still hold, $V(y)$ can be represented by the following equation:

$$V(y) = \kappa_1 V_1[\lambda_1 U_1(y_1) + \lambda_2 U_2(y_2) + \lambda\lambda_1\lambda_2 U_1(y_1)U_2(y_2)] + \kappa_3 V_3(y_3) \quad (10)$$

This follows from theorem (6.1) in Keeney and Raiffa. In equation (10) three trade-off questions have to be asked to determine all the scaling constants: λ_1, λ_2 and one of κ_1 and κ_3.

9.9 Concluding remarks

The aim of this chapter was to derive equity algorithms for QALY based decision making. Two interpretations of QALYs were considered: QALYs as (vNM) utilities and QALYs as measures of health. A justification was provided for aggregating QALYs as consistently scaled vNM utilities over individuals.

It was shown that some of the conditions underlying the common practice of unweighted aggregation of QALYs over individuals are at variance with two types of equity concerns By relaxing some of the conditions underlying the QALY utilitarian model, alternative aggregation procedures were proposed that take into account (some of the) equity considerations. Incorporating equity concerns can be achieved at relatively low cost: a few trade-off questions are sufficient to elicit the preferences of policy makers. Obviously, trade-offs between efficiency and equity considerations are not always easy to make. However, this can be no excuse for not making this trade-off explicit. As argued in this chapter, multi-attribute utility theory can be of great help here.

Appendix 1: Proof of theorem 9.1

Condition *H* is in fact equivalent to the condition of additive independence, which is familiar in multi-attribute utility theory. Additive independence asserts that preferences with respect to lotteries over alternatives depend only on the marginal probability of each outcome occurring and not on the joint probability distribution. Given that it has been assumed that $\succeq_z$ has been defined over a set of (simple) probability distributions, and satisfies the von Neumann Morgenstern axioms, theorem 11.1 in Fishburn (1970)[20] can be applied. According to this theorem:

$$U(x) = \sum_{i=1}^{n} U_i(x_i) \qquad (A\,1)$$

where the $U_i(x_i)$, called additive individual utility functions, are defined from the expected utility of the degenerate lottery that gives outcome x_i with probability *1*. Given the fact that the additive individual utility functions $U_i(x_i)$ are unique up to similar positive linear transformations, it follows that the λ_i's in Harsanyi's theorem are positive. This guarantees positive association between individual and social preferences, one of Arrow's (1951a) conditions.

Imposing condition *A* on top of the other conditions, leads to the QALY utilitarian representation. If $(U_1, U_2, \ldots, U_n)$ is an array of representing additive individual utility functions, then by *condition A* so are $(U_2, U_3, \ldots, U_n, U_1)$, $(U_3, U_4, \ldots, U_1, U_2), \ldots, (U_n, U_1, \ldots, U_{n-1})$. Then $\{(1/n)\Sigma_i U_i, (1/n)\Sigma_i U_i, \ldots, (1/n)\Sigma_i U_i\}$ is representing as well and so, by the uniqueness properties of the U_i, is $(\Sigma_i U_i, \ldots, \Sigma_i U_i)$. This shows that the additive individual utility functions can be chosen identical. Set *U* equal to one of these additive individual utility functions and this gives the desired result.

Continuity follows from the continuity of $\succeq$ and from theorem 3.2 in Maas and Wakker (1994). Additive independence implies utility independence, which in turn implies independence. The structural assumptions made in subsection 5.1 ensure that the other conditions in Maas and Wakker (1994) are fulfilled. Weak order has been assumed. Restricted solvability and the Archimedian axiom follow from the fact that $X_i = \mathbb{R}_+$ and $X = \mathbb{R}_+^n$. $\mathbb{R}_+$ is endowed with the usual Euclidean topology, which is connected and separable. $\mathbb{R}_+^n$ is endowed with the product topology and, by theorem 5.3 in Fishburn (1970), is connected and separable. By the proof of theorem 6.14 in

[20] See also theorem 4 in Fishburn (1965)

Krantz, Luce, Suppes and Tversky (1971), continuity of $\succsim$ with respect to a connected product topology implies restricted solvability and the Archimedean axiom.

Appendix 2: Proof of theorem 9.2

By Maas and Wakker (1994, theorem 3.2) conditions *SE* and *UI1* are equivalent to utility independence for all subsets of $I=\{1,\ldots,n\}$ and U is continuous. Then by theorem 6.1 in Keeney and Raiffa (1976), if $\lambda \neq 0$, $U(x)$ can be written as:

$$\lambda U(x)+1=\prod_{i=1}^{n}\ [\lambda\lambda_i U_i(x_i)+1] \qquad (A2)$$

where $U(x)$ and the U_i are scaled between 0 and 1. Suppose the U_i are scaled according to the algorithm proposed by Gafni and Birch (1991). Then a life in full health has utility 1 for all individuals. If $\lambda=0$, it follows from equation (6.12) in Keeny and Raiffa (1976) that $U(x)$ can be written as equation (A1), which in combination with condition *A* gives QALY-utilitarianism.

Suppose without loss of generality that $(0,0,..,0)$ is the worst social allocation and set $U(0,0,\ldots,0)=0$, which is allowed by free scaling of the utility function. By *condition A*: $(x,0,0,..,0)\sim_z(0,x,0,\ldots,0)\sim_z\ldots\sim_z(0,0,\ldots,0,x)$. Substitute this in equation (A2) to give: $\lambda\lambda_1 U_1(x)=\lambda\lambda_2 U_2(x)=\ldots=\lambda\lambda_n U_n(x)$. Thus all $\lambda_i U_i$ are equal. Set these equal to $U(x_i)$. Rearranging terms gives equation (4).

Denote by $[p,x;(1-p),y]$ a program that gives allocation x with probability p, and allocation y with probability $(1-p)$. By condition *P*, $[0.5,(x,x,0,..,0);\ 0.5,(0,0,..,0)]$ is preferred to $[0.5,(x,0,\ldots,0);0.5,(0,x,0,\ldots,0)]$. Calculating the expected utility of these two programs making use of equation (4) gives:

$$0.5*[(1/\lambda)*(\lambda U(x)+1)^2 - (1/\lambda)]>$$
$$0.5*[(1/\lambda)*(\lambda U(x)+1) - (1/\lambda)]+0.5*[(1/\lambda)*(\lambda U(x)+1) - (1/\lambda)] \qquad (A3)$$

Under the assumption that a QALY is a vNM utility, i.e. $U(x)=x$, eq. (A3) can be rewritten as

$$(1/2\lambda)*(\lambda x+1)^2 - (1/2\lambda)>(1/\lambda)*(\lambda x+1) - (1/\lambda) \qquad (A4)$$

which, after rearranging terms, gives $\lambda\, x/2 > 0$. Given our assumption that x is non-negative and in this particular case cannot equal zero, it follows that $\lambda>0$.

Appendix 3: Proof of theorem 9.3

Given that Y is assumed to be a Cartesian product, that the one-attribute subsets are intervals in the real numbers, that it is implicitly assumed that the decision maker thinks both attributes of $V(Y)$ should influence social preference (i.e. both attributes are essential), and that $\succsim$ on Y is a continuous weak order that satisfies the hexagon condition, by theorem III.4.1. in Wakker (1989) $V(Y)$ can be represented by:

$$V(y) = \kappa_1 V_1(y_1^{ce}) + \kappa_3 V_3(y_3) \qquad (A5)$$

with V and the additive value functions V_i scaled between 0 and 1, continuous and unique up to similar positive linear transformations. The κ_i are scaling constants.

Equivalently equation (A5) can be written as

$$V(y) = \kappa_1 V_1[U(y_1, y_2)] + \kappa_3 V_3(y_3) \qquad (A6)$$

Given that condition *TCI* is equivalent to additive independence for two attributes, theorem 2 in Fishburn (1965) can be applied to give:

$$V(y) = \kappa_1 V_1[\lambda_1 U_1(y_1) + \lambda_2 U_2(y_2)] + \kappa_3 V_3(y_3) \qquad (A7)$$

where the U_i are scaled between 0 and 1, unique up to similar positive linear transformations and are continuous given continuity of V_1.

Finally, it remains to be shown that y_1^{ce} can always be determined. By continuity of the vNM utility function it is possible to find a certainty equivalent for every lottery over y_1 and y_2. Furthermore, given continuity of V_1 $\succsim_A$ restricted to degenerate probability distributions is continuous. By the vNM axioms $\succsim_A$ is a weak order. Finally y_1 and y_2 are elements of intervals in the real numbers. Thus, by lemma III.3.3. in Wakker (1989) $\succsim_A$ satisfies restricted solvability. Suppose (y_1, y_2) denotes the certainty equivalent of a lottery. By restricted solvability if $(a_1, x) \succ_A (y_1, y_2) \succ_A (c_1, x)$ where x denotes the value at which the ex post equity index is held fixed, then there exists (b_1, x) such that $(b_1, x) \sim_A (y_1, y_2)$. If we fix x at the value corresponding to no inequality then there will exist such (a_1, x) and (c_1, x) and thus y_1^{ce} can always be determined.

10

Discussion

This final chapter provides a discussion of the main conclusions drawn in this study. The emphasis will be on discussion rather than on a mere repetition of conclusions that were already stated at the end of the chapters of this thesis. The aim of this final chapter is to bring the various conclusions together in a coherent framework along the lines of the four questions formulated in the introduction and to identify areas for future research.

10.1 QALYs as a utility model

In chapters 2 and 3 I have characterized QALYs as a representation of individual preferences over lotteries on health profiles. Chapter 2 treated the situation where the set of health profiles only contains profiles of a constant quality of life (chronic health states). In chapter 3 the set of health profiles also included health profiles in which quality of life varied over time (i.e. temporary health states are included as well).

It was shown in chapter 2 that in the presence of a condition that is entirely plausible in the medical context, the condition that characterizes the QALY model is *risk neutrality on life years*. Risk neutrality on life years is not a condition that has strong normative appeal. It is hard to conceive of any reason why individuals should behave according to risk neutrality on life years. The implication is that on normative grounds there is no reason to prefer the QALY model. The support for QALYs as a descriptive model is not convincing either. Most empirical studies have rejected risk neutrality on life years [e.g. McNeil et al., 1978; Verhoef et al., 1994; Stiggelbout et al., 1994; Maas and Wakker, 1994]. The exception is Miyamoto and Eraker (1985) who found that risk neutrality on life years holds for the average respondent. However, even in that study risk attitudes varied to a large extent across respondents. The implication of the result presented in chapter 2 is that it may be wise for researchers involved in health care evaluation to test whether individual preferences approximately satisfy risk neutrality on life years when a

QALY model is used to predict choices. Testing for risk neutrality can be done in a relatively straightforward way by asking a limited number of standard gamble questions.

If the preference conditions that characterize the QALY model are found not to hold for a majority of respondents, the use of alternative utility models is recommended. One such model is the general QALY measure proposed by Pliskin et al. (1980) and referred to in the literature as *risk-adjusted QALY model*. In the risk adjusted QALY model life years are adjusted for risk attitude. The two preference conditions that Pliskin et al. use to derive the risk adjusted QALY model are utility independence and constant proportional trade-offs. These conditions were tested in chapter 4 of this thesis. The results of the analysis performed in chapter 4 were relatively favourable to the risk adjusted QALY model. Relatively few respondents satisfied the conditions exactly. However, as is emphasized throughout this thesis, given that respondents are relatively unfamiliar with the techniques of health state utility measurement and with the health states they are asked to evaluate, it is unlikely that they are able to indicate their preferences precisely. Rather they indicate personal confidence intervals. After adjustment for this imprecision in preferences, a majority of respondents satisfied a risk adjusted QALY model. This conclusion held for both health states involved in the study even while they differed substantially in terms of quality of life. It should be kept in mind though that the adjustment for imprecision error applied in chapter 4 was somewhat arbitrary. On the other hand, research aimed at estimating errors of measurement typically indicates confidence intervals that are wider than the ones that were used in chapter 4 [e.g. Torrance, 1976; Rutten-van Mölken et al., 1995].

In chapter 3 I derived that a condition referred to as *additive independence* is the central preference condition in the characterization of the QALY model when quality of life is allowed to vary over time. Additive independence has been tested for chronic health states and has typically been rejected [e.g. Maas and Wakker, 1994]. The results of this empirical research suggest that the QALY model may have to be replaced by a more general model. One such model is the multiplicative model that has been proposed in chapter 3. This model generalizes QALYs by sacrificing the linearity in life years of the QALY model.

Another model that has been proposed as a generalization of the QALY model is the *healthy-years equivalents* (HYEs) [Mehrez and Gafni, 1989]. The development of HYEs has been motivated by reference to empirical violations of the preference conditions characterizing the QALY model for chronic health states. The results of this thesis somewhat moderate claims about the frequency of violation of the conditions underlying the QALY model and thus the need to propose alternative utility based indices. Adjustment for imprecision of preferences

removes the majority of the violations. Gafni, Birch and Mehrez (1993) have argued that HYEs are superior to QALYs because "the HYE approach makes no assumptions about the form of the individual's utility function and thus better reflects the individual's preferences (p.325)." Recent debate in the literature, of which chapter 3 is one contributon, has challenged this claim. The criticism has revealed significant shortcomings of HYEs. The widely accepted conclusion from the recent debate is that the use HYEs in health care analysis is no real improvement over the use of QALYs.

10.2 Methods of health state utility measurement

Chapters 5 and 6 addressed issues of health state utility measurement. In these chapters an attempt was made to answer the second question formulated in the introduction: which of the three methods most commonly used in health state utility measurement (rating scale, time trade-off, and standard gamble) is most consistent with individual preferences? In chapter 5 I examined an inconsistency in utilities elicited by the standard gamble that has been observed before by Llewellyn-Thomas et al. (1982): gambles that are equivalent according to expected utility theory lead to different utilities. It turned out from the analysis presented in chapter 5 that an important reason why this disparity is observed is because individuals do not enter probabilities linearly in the evaluation of gambles, but apply a *probability weighting* function. The weighting function estimated in chapter 5 on the basis of the individual responses turned out to reflect a very pessimistic attitude with respect to the probability of successful treatment. That is, the weighting function reflects a strong aversion to risk.

The reason why such strong risk aversion may have been observed was explained in chapter 5. The standard gamble as it is typically asked in health state utility measurement is a probability equivalence method. Respondents are asked to indicate their preferences in terms of probabilities. From the literature on decision theory it is known that probability equivalence methods typically lead to strong risk averse behaviour, which in turn translates into utility functions that are too concave [Hershey and Schoemaker, 1985]. The way the assorted gamble questions were designed, involved constructing utilities from the responses to two probability equivalence gambles. Constructing utilities from two probability equivalence questions may have led to extremely concave utilities suggesting strong risk averse behaviour. This strong risk averse behaviour in turn is reflected by the very pessimistic weighting function.

The bad performance of the weighting function proposed by Tversky and Kahneman (1992) is another indication that the above phenomenon was indeed at work. The S-shaped weighting function performed well in explaining probability weighting in gambles where the outcomes are amounts of money. Moreover, the parameter estimates of the probability weighting function are remarkably similar across studies [Camerer and Ho, 1994]. The fact that the weighting function estimated in chapter 5 deviates in such a strong way from the S-shaped weighting function suggests that some other phenomenon has produced the results reported in chapter 5. I believe that this other phenomenon is the extreme risk aversion induced by constructing utilities from probability equivalence questions.

The implication of the above conclusion is that responses to standard gamble questions should not be used directly in QALY calculations. The utilities elicited by standard gamble questions are too concave. Rather probabilities elicited in standard gamble questions should be adjusted for probability weighting. It is not clear which probability weighting function should be applied. The results of chapter 5 suggest a highly pessimistic weighting function. However, the good performance of this pessimistic weighting function may merely reflect the process of constructing utilities from probability equivalence questions. The estimation of the probability weighting function for health outcomes should be an important topic for future research. I encourage other researchers to replicate the findings presented in this thesis and to examine the performance of the pessimistic weighting function in other contexts.

The analysis reported in chapter 6 confirms the conclusion that the standard gamble elicits utilities that are too concave. In chapter 6 QALYs elicited on the basis of rating scale quality weights, time trade-off quality weights and standard gamble quality weights were compared with actual choices. It turned out that in comparison with actual choices, QALYs based on standard gamble weights "overvalued" profiles spent in a health state less attractive than full health. This corresponds to overestimating the utility of less than perfect health states. On the other hand, QALYs based on rating scale weights "undervalued" profiles spent in a health state less attractive than full health. This corresponds to underestimating the utility of less than perfect health states.

The analysis of chapter 6 further displayed that QALYs estimated on the basis of the weights elicited by the time trade-off approximated actual choices best. This suggests that the time trade-off method should be the preferred method in health state utility measurement. However, a qualification can be made to this conclusion. It was observed in chapter 6 that discounting decreased the consistency of QALYs based on time trade-off weights. This observation is consistent with a rule of thumb first identified by Stalmeier et al. (1995) and referred to by these

authors as the "proportional heuristic." According to this proportional heuristic individuals answer time trade-off question by determining the number of years in full health as a fixed percentage of the years in the health state to be evaluated. It should be realized that a heuristic that works well in general may not work in any case. A heuristic at best approximates preferences and is not equivalent to preferences. In spite of the above qualification, this thesis has provided support for the use of the time trade-off in health state utility measurement. It will be interesting to see whether the results of chapter 6 can be replicated in other settings involving other health profiles.

The conclusion that emerges from the above discussion (that the time trade-off is approximately right and that the responses to the standard gamble should be adjusted to allow for probability weighting) is somewhat contrary to previous recommendations in health state utility measurement. The widely held belief in health state utility measurement has been that the standard gamble is the norm and that the time trade-off weights should be adjusted upward to allow for risk attitude [Torrance, 1976; Miyamoto and Eraker, 1985; Stiggelbout et al., 1994]. It was hoped that this would remove the observed disparity between time trade-off and standard gamble utilities. However, this will only hold under expected utility theory. The conclusion we derive is that this procedure will not remove the observed disparity. To remove the disparity, the standard gamble should be adjusted downwards to allow for risk attitude reflected by probability weighting.

10.3 Time preference

Chapters 7 and 8 focused on intertemporal preferences for health. These chapters address the third question formulated in the introduction: is the constant rate discounted utility model an appropriate representation of individual intertemporal preferences for health? Chapter 7 contained a theoretical treatment of the constant rate discounted utility model. The preference conditions that characterize the constant rate discounted utility model were identified and it was argued that both on normative grounds and on the basis of earlier empirical research findings doubts can be raised with respect to the validity of the model. Within the class of models that assume intertemporal separability, the condition that distinguishes the constant rate discounted utility model from alternative theories is stationarity. Loewenstein and Prelec (1992) have proposed to use a generalized stationarity condition rather than stationarity in the characterization of intertemporal preferences. This generalized stationarity condition allows Loewenstein and Prelec to derive a more general class of discounted utility models, a limiting case of which is the constant

rate discounted utility model. The generalization of stationarity was motivated by empirical research which showed that for monetary outcomes individuals do not behave according to the predictions of stationarity. In particular, by stationarity preferences between options should be invariant with respect to the passage of time. However, experimental evidence showed that the impact of constant time differences between two outcomes becomes less significant the more remote the outcomes are in time. This phenomenon is referred to as the "common difference effect."

In chapter 8 I presented an experimental test of stationarity versus generalized stationarity. The results of the experiment rejected stationarity. The violation of stationarity was in the direction predicted by the class of generalized discounted utility models. That is, evidence was obtained that individual intertemporal preferences for health were consistent with the common difference effect.

Generalized discounted utility models are not prohibitively more complicated to use than constant rate discounted utility models. Therefore, on the basis of the findings presented in this thesis, I recommend to replace the constant rate discounted utility model by a generalized discounted utility model when the aim is to describe individual intertemporal preferences for health.

The use of a generalized discounted utility model raises problems of its own. First, a method must be found to obtain reliable estimates of the parameters of the model. To date no studies exist to my knowledge that have attempted this task. A problem here is that the estimation of a discount function for health is not as straightforward as the estimation of a discount function for money. The reason is that the set of health states does not have a one-to-one relationship with the set of real numbers. It may be necessary to estimate a discount function for quantity of life, which has a one-to-one relationship with the set of real numbers, first and subsequently examine whether this discount function also explains intertemporal choices for quality of life.

A further problem with generalized discounted utility models based on the common difference effect has been identified in chapter 7. Generalized discounted utility models, like any variable rate discounted utility model, may give rise to dynamically inconsistent behaviour. Such myopic behaviour may not be desirable from a normative point of view.

The class of generalized discounted utility models considered in this thesis retains the assumption of intertemporal separability of preferences. Generalized discounted utility models do not incorporate intertemporal dependency of preferences. The utility of an outcome received at point in time t is not affected by what has occurred in all points in time before t nor by what occurs at all point in

time after t. The sequencing of outcomes is not allowed to affect preferences. In chapter 7 we concluded from a review of published empirical research that sequence effects may have an important impact on intertemporal preferences. For example, Loewenstein and Sicherman (1991) found that workers generally prefer increasing wage profiles to decreasing wage profiles of the same expected value. Such sequence effects are also likely to be important in health decision making. Phenomena like maximal endurable time and coping can only be explained by taking account of intertemporal dependency of preferences. Using models that assume intertemporal separability of preferences seems more suitable to model short-range decisions than long-range decisions. However, decisions with respect to health generally fall in the latter category.

Several suggestions as to the possibility of modelling sequence effects were considered in chapter 7. One possibility that was considered there is the application in intertemporal decision making of rank dependent models that have been successful in modelling decisions under uncertainty (see also chapter 5). Using an S-shaped weighting (discounting) function in the context of intertemporal choice implies that outcomes at the beginning and at the end of sequences are overweighted. The empirical evidence that is available to date does indeed indicate the importance of the outcomes occurring at the beginning and at the end of sequences. S-shaped weighting also emphasizes the importance of the present relative to the future and therefore incorporates what Prelec and Loewenstein (1991) refer to as "immediacy effects." Immediacy effects are comparable to the effect of certain outcomes in choice under uncertainty. In choice under uncertainty certain outcomes are generally overweighted relative to uncertain outcomes. Similarly, in intertemporal choice the present is generally overweighted relative to the future.

Rank dependent utility theory can incorporate effects that have been shown to be important influences on intertemporal decision making. Further research into its applicability in modelling intertemporal preferences for health promises to be worthwhile.

10.4 Equity principles

Chapter 9 focused on the use of QALYs in social decision making. The aim of chapter 9 was to provide an answer to the fourth question formulated in the introduction: what conditions have to be imposed on the social preference relation to ensure that it can be characterized by unweighted aggregation (QALY utilitarianism)? Related to this question is the problem how equity principles can be incorporated in cost utility analysis. Before these problems could be addressed a

rationale had to be provided for aggregating QALYs over individuals. The question of the appropriateness of aggregating utilities has a long history in economics. Building on recent research, I have made an attempt to argue why aggregation of von Neumann Morgenstern utilities may be allowed.

The analysis presented in chapter 9 showed that QALY utilitarianism excludes two types of equity concern: concern for the final distribution of QALYs (ex post equity)[1] and concern for the fairness of the QALY allocation process (ex ante equity). Utility indices that incorporate these types of equity concern were derived. The use of these alternative utility indices does not prohibitively complicate the calculation of the social benefits of health care programs. However, they require the elicitation of preferences from policy makers with respect to the desirable rate of trade-off between equity and efficiency (expressed in terms of the total number of QALYs gained) aspects of health care programs. The tools of multi-attribute utility theory can be helpful to gain insight in this rate of trade-off. The application of multi-attribute utility theory in this context does not seem to be more complicated than in other contexts where the theory has been successfully applied. As long as no insight exists in the societal trade-off between equity and efficiency, it is recommended to use the utility models developed in chapter 9 in a sensitivity analysis. Several values for the equity efficiency trade-off can be hypothesized and the sensitivity of the results for these different specifications can be examined.

The models specified in chapter 9 still depend on preference conditions that may not be realistic in every setting. The main motivation for the utility indices presented in chapter 9 was simplicity. A discrepancy exists in health economics and medical decision making between the number of contributions drawing attention to the importance of incorporating distributional considerations in QALY based decision making and the number of contributions that have actually attempted to propose ways to incorporate these distributional concerns. The aim of chapter 9 was to show that distributional concerns can be incorporated in QALY based decision making without making the analysis extremely complicated. To achieve this aim several assumptions had to be retained. Relaxing these assumptions will make the utility models at the same time more realistic and more complicated.

---

[1] Note that in this thesis I have not considered a concern for the final distribution of QALYs which is motivated by feelings of altruism. Such a concern is strictly speaking an efficiency issue: an individual will favour providing care to the less well off as long as the marginal utility of providing such care exceeds the marginal costs. For a treatment of such externalities in cost utility analysis see Labelle and Hurley (1992).

One way to generalize the models proposed in chapter 9 is to translate rank dependent utility models to the context of social decision making over allocations. The rank dependent utility model has already been successfully applied in the social welfare literature [Weymark, 1981; Ebert, 1988; Ben Porath and Gilboa, 1994]. Rank dependent utility models amount to giving different weights to different outcomes. In the present case, the weighting function applies to the number of QALYs received by different individuals. For example, the application of a pessimistic weighting function implies that individuals who are relatively well off in terms of QALYs gained are underweighted in comparison with individuals who are relatively worse off in tems of QALYs gained. The equity principles proposed in chapter 9 are based on similar principles, but require more restrictive assumptions. The applicability of rank dependent utility theory in the context of social decision making over QALY allocations is worthy of future research.

10.5 Epilogue

The subject matter of this thesis arose out of existing confusion in the literature over the role in health care analysis of health utility indices in general and QALYs in particular. This confusion was mainly caused by the existing divergence in health economics between the large number of practical applications of utility based decision making and the number of methodological issues that had been solved. The aim of this thesis was to provide a contribution to remove this divergence. Without claiming comprehensiveness of treatment, I believe that this thesis has helped to clarify the waters to some extent. Various theoretical results have been derived that give health utility indices a firmer foundation in utility theory. These theoretical results help to assess the normative and descriptive validity of health utility indices. Moreover, empirical research has provided insight in the descriptive validity of various models commonly used in utility based decision making in health.

The overall message of this thesis is moderately positive for QALY based decision making. Even though in the theoretical analyses presented in this thesis I generally concluded that the preference conditions underlying QALY based decision making were restrictive, a reasonable degree of support was found for QALY based decision making in the empirical analyses.This is no argument for complacency. Various issues remain to be tackled. However, the results of this thesis are at least to some extent reassuring with respect to the direction health care analysis has taken over the past two decades.

References

Abdelbasit, K.M. and R.L. Plackett, 1983, Experimental design for binary data, Journal of the American Statistical Association 78, 90-98.

Allais, M., Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Américaine, Econometrica 21, 503-546.

Altman, D.G., 1991, Practical statistics for medical research (Chapman & Hall, London).

Amemiya, T., 1994, Studies in econometric theory (Edward Elgar, Aldershot)

Arrow, K.J., 1950, A difficulty in the concept of social welfare, Journal of Political Economy 58, 328-346.

Arrow, K.J., 1951a, Social choice and individual values (Wiley, New York). (2nd ed., 1963)

Arrow, K.J., 1951b, Alternative approaches to the theory of choice in risk-taking situations, Econometrica 19, 404-437.

Atkinson, A.B., 1970, On the measurement of inequality, Journal of Economic Theory 2, 244- 263.

Ayer, M., H.D. Brunk, G.M. Ewing and E. Silverman, 1955, An empirical distribuion function for sampling with incomplete information, Annals of Mathematical Statistics 26, 641-647.

Bakker C., M. Rutten-van Mölken, E. van Doorslaer, K. Bennett and Sj. van der Linden, 1994, Feasibility of utility assessment by rating scale and standard gamble in ankylosing spondylitis or fibromyalgia, Journal of Rheumatology 21, 269-274.

Ben Porath, E. and I. Gilboa, 1994, Linear measures, the Gini index and the income-equality trade-off, Journal of Economic Theory 64, 443-467.

Ben-Zion, U., Gafni, A., 1983, Evaluation of public investment in health care: Is the risk irrelevant?, Journal of Health Economics 2, 161-165.

Blackorby, C., D. Primont and R.R. Russell, 1978, Duality, separability, and functional structure: Theory and economic applications (Elsevier, New York).

Blau, J. H., 1957, The existence of a social welfare function, Econometrica 25, 302-313.

Bleichrodt, H., 1995, QALYs and HYEs: under what conditions are they equivalent? Journal of Health Economics 14, 17-37.

Bleichrodt, H. and A. Gafni, 1995, Time preference, the discounted utility model and health, Journal of Health Economics (in press).

Broome, J., 1982, Equity in risk bearing, Operations Research 30, 412-414.

Broome, J., 1985, The economic value of life, Economica 52, 281-294.

Broome, J., 1988, Goodness, fairness and QALYs, in: M. Bell and S. Mendus, eds. Philosophy and medical welfare (Cambridge Univ. Press, Cambridge), 57-73.

Broome, J., 1993, Qalys, Journal of Public Economics 50, 149-167.

Buckingham, K., 1993, A note on HYE, Journal of Health Economics 12, 301-309.

Cairns, J., 1994, Valuing future benefits, Health Economics 3, 221-229.

Camerer, C. and T. Ho, 1994, Violations of the betweenness axioms and non-linearity in probability, Journal of Risk and Uncertainty 8, 167-196.

Constantinides, G.M., 1990, Habit formation: A resolution of the equity premium puzzle, Journal of Political Economy 98, 519-543.

Culyer, A.J., 1989, The normative economics of health care finance and provision, Oxford Review of Economic Policy 5, 34-58.

Culyer, A.J. and A. Wagstaff, 1993, QALYs versus HYEs, Journal of Health Economics 12, 311-323.

d'Aspremont, C. and L. Gevers, 1977, Equity and the informational basis of collective choice, Review of Economic Studies 44, 199-208.

Davidson, R. and J.G. MacKinnon, 1984, Convenient specification tests for logit and probit models, Journal of Econometrics 25, 241-262.

Debreu, G., 1954, Representation of a preference ordering by a numerical function, in: R.M. Thrall, C.H. Coombs and R.L. Davis, eds. Decision processes (Wiley, New York), 159-165.

Debreu, G., 1960, Topological methods in cardinal utility theory, in: K.J. Arrow, S. Karlin and P. Suppes, eds. Mathematical methods in the social sciences (Stanford University Press), 16-26.

Debreu, G., 1964, Continuity properties of Paretian utility, International Economic Review 5, 285-293.

Deschamps, R. and L. Gevers, 1978, Leximin and utilitarian rules: A joint characterization, Journal of Economic Theory 17, 143-163.

Diamond, P.A., 1967, Cardinal welfare individualistic ethics and interpersonal comparisons of utility: Comment, Journal of Political Economy 75, 765-766.

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

Dyer, J.S. and R.K. Sarin, 1979, Measurable multiattribute value functions, Operations Research 27, 810-822.

Ebert, U., 1988, Measurement of inequality: An attempt at unification and generalization, Social Choice and Welfare 5, 147-169.

Ellsberg, D., 1954, Classic and current notions of "measurable utility", Economic Journal 64, 528-556.

Eriksen, S. and L.R. Keller, 1993, A multiattribute utility function approach to weighing the risks and benefits of pharmaceutical agents, Medical Decision Making 13, 118-125.

EuroQol Group, 1990, EuroQol: a new facility for the measurement of health-related quality of life, Health Policy 16, 199-208.

Farquhar P.H., 1984, Utility assessment methods, Management Science 30, 1283-1300.

Fishburn, P.C., 1965, Independence in utility theory with whole product sets, Operations Research 13, 28-45.

Fishburn, P.C., 1970, Utility theory for decision making (Wiley, New York).

Fishburn, P.C., 1984, Equity axioms for public risk, Operations Research 32, 901-908.

Fishburn, P.C., 1989, Retrospective on the utility theory of von Neumann and Morgenstern, Journal of Risk and Uncertainty 2, 127-158.

Fishburn, P.C. and A. Rubinstein, 1982, Time preference, International Economic Review 23, 677-694.

Fishburn, P.C. and P. Straffin, 1989, Equity considerations in public risk evaluation, Operations Research 37, 229-239.

Fleming, M., A cardinal concept of welfare, Quarterly Journal of Economics 66, 366-384.

Frank, R. and R. Hutchens, 1993, Wages, seniority, and the demand for rising consumption profiles, Journal of Economic Behavior and Organization 21, 251-276.

Gafni, A., 1994, The standard gamble method: What is being measured and how it is intrpreted, Health Services Research, 29, 207-224.

Gafni, A., 1995, Time in health: Can we measure individual's pure time preference?, Medical Decision Making, 15, 31-37.

Gafni, A. and S. Birch, 1991, Equity considerations in utility-based measures of health outcomes in economic appraisals: an adjustment algorithm, Journal of Health Economics 10, 329-342.

Gafni, A., S. Birch and A. Mehrez, 1993, Economics, health and health economics: HYEs versus QALYs, Journal of Health Economics 12, 325-339.

Gafni, A. and G.W. Torrance, 1984, Risk attitude and time preference in health, Management Science 30, 440-451.

Gafni, A. and C.J. Zylak, 1990, Ionic versus non-ionic contrast media: a burden or a bargain? Canadian Medical Association Journal 140, 475-478.

Gilboa, I., 1989, Expectation and variation in multi-period decisions, Econometrica 57, 1153-1169.

Greene W., 1993, Econometric analysis, 2nd ed. (MacMillan, New York).

Grossman, M., 1972, The demand for health: a theoretical and empirical investigation, National Bureau of Economic Research, New York.

Hammond, P., 1976, Changing tastes and coherent dynamic choice, Review of Economic Studies 43, 159-173.

Hanemann, W.M., 1991, Willingness to pay and willingness to accept: how much can they differ?, American Economic Review 81, 635-647.

Harless, D.W. and C. Camerer, 1994, The predictive utility of generalized expected utility theories, Econometrica 62, 1251-1289.

Harris, J., 1988, More and better justice, in: M. Bell and S. Mendus, eds. Philosophy and medical welfare (Cambridge Univ. Press, Cambridge), 75-96.

Harsanyi, J.C., 1955, Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility, Journal of Political Economy 63, 309-321.

Harsanyi, J.C., 1975, Non-linear social welfare functions, or do welfare economists have a special exemption from Bayesian rationality, Theory and Decision 6, 311-332.

Harsanyi, J.C., 1987, von Neumann Morgenstern utilities, risk taking and welfare, in: G.R. Feiwel, ed., Arrow and the ascent of modern economic theory (Macmillan, London).

Harvey, C.M., 1994, The reasonableness of non-constant discounting, Journal of Public Economics 53, 31-51.

Hausman, J.A., ed., 1993, Contingent valuation: a critical assessment (Elsevier, Amsterdam).

Hershey, J.C. and P.J.H. Schoemaker, 1985, Probability versus certainty equivalence methods in utility measurement: are they equivalent? Management Science 31, 1213-1231.

Hildreth, C., 1953, Alternative conditions for social ordering, Econometrica 21, 81-89.

Hornberger J.C., D.A. Redelmeier and J. Peterson, 1992, Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis, Journal of Clinical Epidemiology 45, 505-512.

Jensen, N.E., 1967, An introduction to Bernoullian utility theory: I. Utility functions, Swedish Journal of Economics 69, 163-183.

Johannesson M., 1995, Quality-adjusted life-years versus healthy-years equivalents: a comment, Journal of Health Economics 14, 9-16.

Johannesson, M., J.S. Pliskin and M.C. Weinstein, 1993, Are healthy years equivalents an improvement over quality adjusted life years? Medical Decision Making 13, 281- 286.

Johannesson, M., J.S. Pliskin and M.C. Weinstein, 1994, A note on QALYs, time trade-off, and discounting, Medical Decision Making 14, 188-193.

Kahneman, D. and R. Thaler, 1991, Economic analysis and the psychology of utility: Applications to compensation policy, American Economic Review Proceedings 81, 341-346.

Kahnemann, D. and A. Tversky,1979, Prospect theory: an analysis of decision-making under risk, Econometrica 47, 263-291.

Karni E. and Z. Safra, 1990, Rank-dependent probabilities, Economic Journal 100, 487-495.

Karni, E. and D. Schmeidler, 1991, Utility theory with uncertainty, in: W. Hildenbrand and H. Sonnenschein, eds. Handbook of mathematical economics, Vol. 4 (North-Holland, Amsterdam), 1763-1831.

Keeney, R., 1980, Equity and public risk, Operations Research 28, 527-533.

Keeney, R. and H. Raiffa, 1976, Decisions with multiple objectives (Wiley, New York).

Koopmans, T.C., 1960, Stationary ordinal utility and impatience, Econometrica 28, 287-309.

Koopmans, T.C., 1972, Representation of preference orderings over time, in: C.B. McGuire and R.Radner, eds. Decision and organization (North-Holland, Amsterdam), 79-100.

Koopmans, T.C., P.A. Diamond and R.E. Williamson, 1964, Stationary utility and time perspective, Econometrica 32, 82-100.

Krahn, M. and A. Gafni, 1993, Discounting in the economic evaluation of health care interventions, Medical Care 31, 403-418.

Krantz, D.H., R.D. Luce, P. Suppes and A. Tversky, 1971, Foundations of measurement, vol. 1 (Academic Press, New York).

Kriström, B., 1990a, Valuing environmental benefits using the contingent valuation method: An econometric analysis, Ph.D. thesis, University of Umeå, Umeå Ecnomic Studies No. 219.

Kriström, B., 1990b, A non-parametric approach to the estimation of welfare measures in discrete response valuation studies, Land Economics 66, 135-139.

Labelle, R.J. and J.E. Hurley, 1992, Implications of basing health care resource allocations on cost utility analysis in the presence of externalities, Journal of Health Economics 11, 259-277.

Landis, J.R. and Koch G.G., The measurement of observer agreement for categorical data, Biometrics, 33, 159-174.

Lind, R.C., 1990, Reassessing the government's discount rate policy in the light of new theory and data in a world economy with a high degree of capital mobility, Journal of Environmental Economics and Management 18, S8-28.

Lipscomb, J., 1989, Time preference for health in cost-effectiveness analysis, Medical Care 27, S233-S253.

Llewellyn-Thomas, H., H.J. Sutherland, R. Tibshirani, A. Ciampi, J.E. Till and N.F. Boyd, 1982, The measurement of patients' values in medicine, Medical Decision Making 2, 449-462.

Llewellyn-Thomas H., et al., 1984, Describing health states: methodological issues in obtaining values for health states, Medical Care 22, 543-552.

Lockwood, M., 1988, Quality of life and resource allocation, in: M. Bell and S. Mendus, eds. Philosophy and medical welfare (Cambridge Univ. Press, Cambridge), 33-55.

Loewenstein, G., 1987, Anticipation and the value of delayed consumption, Economic Journal 97, 666-684.

Loewenstein, G., 1988, Frames of mind in intertemporal choice, Management Science 34, 200-214.

Loewenstein, G. and D. Prelec, 1991, Negative time preference, American Economic Review Proceedings, 81, 347-352.

Loewenstein, G. and D. Prelec, 1992, Anomalies in intertemporal choice: Evidence and an interpretation, Quarterly Journal of Economics 107, 573-597.

Loewenstein, G. and D. Prelec, 1993, Preferences for sequences of outcomes, Psychological Review 100, 91-108.

Loewenstein, G. and N. Sicherman, 1991, Do workers prefer increasing wage profiles?, Journal of Labor Economics 9, 67-84.

Loewenstein, G. and R. Thaler, 1989, Anomalies: Intertemporal choice, Journal of Economic Perspectives 3, 181-193.

Loomes, G., 1995, The myth of the HYE, Journal of Health Economics 14, 1-7.

Loomes, G. and L. McKenzie, 1989, The use of QALYs in health care decision making, Social Science and Medicine 28, 299-308.

Luce, R.D., B.A. Mellers and S-J Chang, 1993, Is choice the correct primitive? On using certainty equivalents and reference levels to predict choices among gambles, Journal of Risk and Uncertainty 6, 115-143.

Luce, R.D. and H. Raiffa, 1957, Games and decisions (Wiley, New York).

Maas A. and P.P. Wakker, 1994, Additive conjoint measurement for multiattribute utility, Journal of Mathematical Psychology 38, 86-101.

Mackeigan, L.D., L.N. Lawson, J.R. Drangalis, J.L. Bootman and L.R. Burns, 1993, Time preference for health gains versus health losses, Pharmaco Economics, 3, 374-386.

McCloskey, D.N., 1983, The rhetoric of economics, Journal of Economic Literature, 21, 481-517.

McNeil B.J., R. Weichselbaum and S.G. Pauker, 1978, Fallacy of the five-year survival in lung cancer, New England Journal of Medicine 299, 1397-1401.

McNeil, B.J., R. Weichselbaum and S.G. Pauker, 1981, Tradeoffs between quality and quantity of life in laryngeal cancer, New England Journal of Medicine 305, 982-987.

Mehrez, A. and A. Gafni, 1989, Quality-adjusted life-years, utility theory and healthy-years equivalents, Medical Decision Making 9, 142-149.

Mehrez, A. and A. Gafni, 1991, The healthy-years equivalents: how to measure them using the standard gamble approach, Medical Decision Making 11, 140-146.

Miyamoto J.M. and S.A. Eraker, 1985, Parameter estimates for a QALY utility model, Medical Decision Making 5, 191-213.

Miyamoto, J.M. and S.A. Eraker, 1988, A multiplicative model of the utility of survival duration and health quality, Journal of Experimental Psychology 117, 3-20.

Mooney, G. and J.A. Olsen, 1991, QALYs: where next? in: A. McGuire, A.K. Mayhew and P. Fenn, eds. The economics of alternative systems of health care finance and delivery (Oxford University Press), 120-140.

Nord, E., 1994, The QALY-a measure of social value rather than individual utility?, Health Economics 3, 89-93.

Nord , E., 1995, The person-trade-off approach to valuing health care programs, Medical Decision Making, 15, 201-208.

Olsen, J.A., 1993a, Time preference for health gains: An empirical investigation, Health Economics 2, 257-265.

Olsen, J.A., 1993b, On what basis should health be discounted, Journal of Health Economics 12, 39-53

Olson, M. and M.J. Bailey, 1981, Positive time preference, Journal of Political Economy 89, 1-25.

Pliskin J.S., D.S. Shepard and M.C. Weinstein, 1980, Utility functions for life years and health status, Operations Research 28, 206-24.

Pollak, R.A., 1970, Habit formation and dynamic demand functions, Journal of Political Economy 78, 745-763.

Prelec, D. and G. Loewenstein, 1991, Decision making over time and under uncertainty: A common approach, Management Science 37, 770-786.

Quiggin J., 1982, A theory of anticipated utility, Journal of Economic Behaviour and Organization 3, 323-343.

Quinn R.J.,1981, The effect of measurement method on preference scales in a medical decision-making context, Medical Decision Making 1, 431.

Read J.L., R.J. Quinn, D.M. Berwick, H.V. Fineberg and M.C. Weinstein, 1984, Preferences for health outcomes: comparison of assessment methods, Medical Decision Making 4, 315-329.

Redelmeier, D.A. and D.N. Heller, 1993, Time preference in medical decision making and cost effectiveness analysis, Medical Decision Making 13, 212-217.

Richardson, J., 1994, Cost utility analysis: What should be measured? Social Science and Medicine , 7-20.

Rutten-van Mölken, M., C. Bakker, E. van Doorslaer and Sj. van der Linden, 1995, Methodological issues in utility measurement, Medical Care (in press).

Sackett, D.L. and G.W. Torrance, 1978, The utility of different health states as perceived by the general public, Journal of Chronical Diseases 31, 697-704.

Samuelson, P.A., 1952, Probability, utility and the independence axiom, Econometrica 20, 670-678.

Savage, L.J., 1954, The foundations of statistics (Wiley, New York).

Schoemaker, P.J.H., 1982, The expected utility model: Its variants, purposes, evidence and limitations, Journal of Economic Literature 20, 529-563.

Schoemaker, P.J.H., 1993, Determinants of risk-taking: Behavioral and economic views, Journal of Risk and Uncertainty 6, 49-73.

Scitovsky, T., 1976, The joyless economy (Oxford University Press).

Sen, A.K., 1970, Collective choice and social welfare (Holden-Day, San Francisco).

Sen, A.K., 1973, On economic inequality (Clarendon Press, Oxford).

Sen, A.K., 1976, Welfare inequalities and Rawlsian axiomatics, Theory and Decision 7, 243-262.

Sen, A.K., 1977, On weights and measures: Informational constraints in social welfare analysis, Econometrica 45, 1539-1572.

Sen, A.K., 1979, Personal utilities and public judgements: Or what's wrong with welfare economics ?, Economic Journal 89, 537-558.

Sen, A.K., 1986, Social choice theory, in: K.J. Arrow and M. Intriligator, eds., Handbook of Mathematical Economics, vol. III (North Holland, Amsterdam), 1073-1182.

Smith, A., 1987, Qualms about QALYs, The Lancet i, 1134.

Stalmeier P.F.M., T.G.G. Bezembinder and I.J. Unic, 1995 Proportional heuristics in time trade off and conjoint measurement, Medical Decision Making (in press).

Stiggelbout A.M., G.M. Kiebert, J. Kievit, J.W.H. Leer, G. Stoter and J.C.J.M. de Haes, 1994, Utility assessment in cancer patients: adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores, Medical Decision Making 14, 82-90.

Strotz, R.H., 1956, Myopia and inconsistency in dynamic utility maximization, Review of Economic Studies 23, 165-180.

Sutherland, H.U., H. Llewellyn-Thomas, N.F. Boyd and J.E. Till, 1982, Attitudes toward quality of survival: the concept of "maximal endurable time", Medical Decision Making 2, 299-309.

Torrance G.W., 1976, Social preferences for health states: an empirical evaluation of three measurement techniques, Socioeconomic Planning 10, 129-136.

Torrance, G.W., 1986, Measurement of health state utilities for economic appraisal: a review, Journal of Health Economics 5, 1-30.

Torrance, G.W., M.H. Boyle and P.H. Horwood, 1982, Application of multi-attribute utility theory to measure social preferences for health states, Operations Research 30, 1043-1069.

Torrance, G.W. and D. Feeny, 1989, Utilities and quality-adjusted life years, International Journal of Technology Assessment in Health Care 5, 559-575.

Torrance, G.W., W.H. Thomas, and D.L. Sackett, 1972, A utility maximization model for evaluation of health care programmes. Health Services Research 7, 118-33.

Torrance, G.W., Y. Zhang, D. Feeny, W. Furlong and R. Barr, 1992, Multi-attribute preference functions for a comprehensive health status classification system, McMaster University CHEPA working paper no. 92-18.

Tversky, A. and D. Kahneman, 1991, Loss aversion in riskless choices: a reference dependent model, Quarterly Journal of Economics 106, 1039-1061.

Tversky, A. and D. Kahneman, 1992, Advances in prospect theory: cumulative representation of uncertainty, Journal of Risk and Uncertainty 5, 297-323.

Tversky, A. and P.P. Wakker, 1995, Risk attitudes and decision weights, Econometrica (in press)

Ulph, A., 1982, The role of ex-ante and ex-post decisions in the valuation of life, Journal of Public Economics 18, 265-276.

Verhoef L.C.G., A.F.J. de Haan and W.A.J. van Daal, 1994, Risk attitude in gambles with years of life: empirical support for prospect theory, Medical Decision Making 14, 194-200.
Vind, K., Additive utility functions and other special functions in economic theory, Discussion paper 90-21, Institute of Economics, University of Copenhagen.

Viscusi, W.K. and W.N. Evans, 1990, Utility functions that depend on health status: estimates and economic implications, American Economic Review 80, 353-374.

von Neumann J. and O. Morgenstern, 1944, Theory of Games and Economic Behavior (Princeton University Press, Princeton)

Wagstaff, A., 1991, QALYs and the equity-efficiency trade-off, Journal of Health Economics 10, 21-41.

Wakker, P.P., 1984, Cardinal coordinate independence for expected utility, Journal of Mathematical Psychology 28, 110-117.

Wakker, P.P., 1989, Additive representations of preferences: A new foundation of decision analysis (Kluwer, Dordrecht).

Wakker, P.P., 1994, Separating marginal utility and risk aversion, Theory and Decision 36, 1-44.

Wakker P.P. and A. Stiggelbout, 1995, Explaining distortions in utility elicitation through the rank-dependent model for risky choices, Medical Decision Making 15, 180-186.

Tversky, A. and P.P. Wakker, 1995, Risk attitudes and decision weights, Econometrica (in press)

Weinstein. M.C., 1993, Time-preference studies in the health care context, Medical Decision Making 13, 218-219.

Weinstein, M.C., H.C. Fineberg et al., 1980, Clinical decision analysis (W.B. Saunders, Philadelphia PA).

Weymark, J., 1981, Generalized Gini inequality indices, Mathematical Social Sciences 1, 409-430.

Williams, A., 1981, Welfare economics and health status measurement, in: J. van der Gaag and M. Perlman, eds., Health, economics and health economics (North Holland, Amsterdam), 271-281.

Williams, A., 1993, Priorities and research strategy in health economics for the 1990s, Health Economics 2, 295-302.

Wolfson, A., A. Sinclair, C. Bombardier and A. McGeer,1982, Preference measures for functional status in stroke patients: interrater and intertechnique comparisons, in: R.A. Kane and R.L. Kane, eds. Values and long term care Lexington MA), 191-214.

Yatchew, A. and Z. Griliches, 1985, Specification errors in probit models, Review of Economics and Statistics 66, 134-139.

Samenvatting

Toepassingen van nutstheorie in de economische evaluatie van gezondheidszorg

1. Inleiding

Dit proefschrift bestudeert de toepasbaarheid van "quality-adjusted life years" (QALYs) en andere op nutstheorie gebaseerde uitkomstmaten in medisch besliskundig en gezondheidseconomisch onderzoek. De voornaamste conclusie van het proefschrift is dat op nut gebaseerde uitkomstmaten (utiliteitsmaten) het keuzegedrag met betrekking tot gezondheid betrouwbaarder modelleren dan vaak is aangenomen.

Een voorbeeld kan de toepassing van utiliteitsmaten in gezondheidsonderzoek illustreren. Stel een individu moet een keuze maken uit twee behandelingen voor een ernstige huidziekte. De eerste behandeling verbetert de huidziekte van een ernstige vorm tot een lichte vorm gedurende een periode van 10 jaar. De tweede behandeling levert de eerste twee jaar geen verbetering op, maar na twee jaar verdwijnt de huidziekte gedurende een periode van 8 jaar. Na 10 jaar zijn beide behandelingen uitgewerkt en verslechtert de huidziekte weer tot ernstig. De kosten van beide behandelingen zijn op betrouwbare wijze gemeten en zijn uitgedrukt in geld. De vraag is nu hoe tussen beide behandelingen een keuze gemaakt kan worden. Een mogelijkheid is eenvoudigweg die behandeling te kiezen die het goedkoopst is. Een dergelijke vorm van analyse wordt omschreven met de term "kosten-minimalisatie". Kosten minimalisatie houdt geen rekening met het feit dat de twee behandelingen verschillen in de gezondheidsuitkomsten die zij genereren. Om tot een goede afweging van kosten en baten (gezondheidswinst) te komen is het noodzakelijk om de baten van gezondheidszorgprogramma's in een gemeenschappelijke eenheid uit te drukken. Een eerste mogelijkheid is om de baten van een gezondheidsprogramma uit te drukken in gewonnen levensjaren. Deze uitkomstmaat gaat echter voorbij aan het feit dat een belangrijk doel van veel

gezondheidsprogramma's niet het verlengen van leven is, maar het verhogen van de kwaliteit waarin levensjaren worden doorgebracht. In bovenstaand voorbeeld leidt behandeling niet tot extra levensjaren en zouden de baten van beide behandelingen volgens de uitkomstmaat "gewonnen levensjaren" dus gelijk zijn aan nul. Beide behandelingen leiden echter tot een verhoogde kwaliteit van leven en de conclusie dat zij geen baten opleveren is daarom niet juist. Er bestaat behoefte aan een methode die in staat is de meervoudige dimensie van gezondheidsbaten te reflecteren.

Omdat de kosten in geld zijn uitgedrukt lijkt geld ook de meest logische eenheid om de baten van gezondheidszorgprogramma's in uit te drukken. De analysevorm waarin zowel kosten als baten in geld worden uitgedrukt staat bekend onder de naam "kosten-baten analyse". In kosten-baten analyse wordt een keuze gemaakt tussen twee programma's op basis van het criterium welk programma het grootste overschot van baten over kosten oplevert. Er bestaan in principe twee methodes om de baten van een gezondheidszorgprogramma in geld uit te drukken. De eerste methode vraagt individuen naar hun bereidheid tot betalen ("willingness to pay") voor verbeteringen in gezondheid (of voor het voorkomen van verslechteringen in gezondheid). In bovenstaand voorbeeld zouden de monetaire baten van gezondheidsprogramma 1 bepaald kunnen worden door te vragen welk bedrag een individu maximaal bereid is te betalen voor een verbetering van zijn gezondheid van een ernstige vorm van huidziekte tot een lichte vorm van huidziekte. De tweede methode vraagt individuen hoeveel geld zij bereid zijn te accepteren ("willingness to accept") voor een verslechtering van hun gezondheid. In bovenstaand voorbeeld kunnen de monetaire baten van gezondheidsprogramma 1 worden bepaald door een individu te vragen hoeveel geld hij minimaal bereid is te accepteren voor een verslechtering van zijn gezondheid van een lichte vorm van huidziekte tot een ernstige vorm van huidziekte.

Theoretisch zouden de twee methodes ongeveer gelijke resultaten moeten geven.[1] Empirisch onderzoek heeft echter aangetoond dat de uitkomsten van de twee methodes sterk verschillen. Het minimum bedrag dat individuen bereid zijn te accepteren is in het algemeen een veelvoud van het maximum bedrag dat zij bereid zijn te betalen voor een gegeven verandering in gezondheid. Naast deze discrepantie heeft empirisch onderzoek ook het bestaan van andere verstorende factoren bij de bepaling van de monetaire waarden van gezondheidstoestanden aangetoond. Een overzicht van deze verstorende factoren geeft het boek van Hausman (1993).

---

[1] Een beperkte afwijking is mogelijk als gevolg van het optreden van substitutie effecten [zie Hanemann (1991)]

De problemen met het bepalen van monetaire waarden voor gezondheidstoestanden hebben geleid tot de ontwikkeling van een alternatieve analysemethode: kosten-utiliteitsanalyse. In een kosten-utiliteitsanalyse worden de baten van een gezondheidszorgprogramma niet in monetaire eenheden, maar in eenheden van nut uitgedrukt. De meest gebruikelijke nutsindex in kosten-utiliteitsonderzoek is de zogenaamde QALY. QALYs worden berekend door levensjaren te aan te passen voor de kwaliteit van leven waarin deze jaren worden doorgebracht. De aanpassing voor kwaliteit van leven vindt plaats door utiliteiten toe te kennen aan gezondheidstoestanden. Stel dat we in bovenstaand voorbeeld er in geslaagd zijn om te bepalen dat de utiliteit van een ernstige vorm van huidziekte gelijk is aan 0.50. Dit schrijven we als *u(ernstige huidziekte) = 0.50.* Verder hebben we bepaald dat *u(lichte huidziekte) = 0.80* en *u(geen huidziekte)= 1.* Behandeling 1 levert dan *10*0.8=8* QALYs op en behandeling 2 levert *2*0.50+8*1=9* QALYs op. De kosten van de programma's buiten beschouwing latend, zou op basis van het criterium "maximaliseer het aantal QALYs" gekozen moeten worden voor behandeling 2.

Laten we het bovenstaande iets formeler uitdrukken. Stel een bepaalde behandeling resulteert in T jaren die in kwaliteit van leven q_t worden doorgebracht. Het symbool q_t staat voor "kwaliteit van leven in periode t". Het aantal QALYs dat deze behandeling oplevert, wordt berekend als

$$\sum_{t=1}^{T} u(q_t) \qquad (1)$$

De functie $u(q_t)$ kan worden geïnterpreteerd als een nutsfunctie over kwaliteit van leven. Deze functie kent utiliteiten aan gezondheidstoestanden toe. In bovenstaand voorbeeld *u(ernstige huidziekte)=0.50.*

De drie meest gebruikte methodes in gezondheidsonderzoek om utiliteiten voor gezondheidstoestanden te bepalen zijn de ratioschaal, de "time trade-off" en de "standard gamble". Een probleem is dat de methodes resulteren in verschillende utiliteiten. Deze discrepantie in utiliteiten kan als consequentie hebben dat de uitkomst van een kosten-utiliteitsanalyse varieert met de gehanteerde methode om de utiliteiten te bepalen. Wanneer in bovenstaand voorbeeld zou gelden dat volgens een andere methode *u(ernstige huidziekte)=0.45* en *u(lichte huidziekte)=0.90*, dan zou behandeling 1 geprefereerd worden. Dergelijke verschillen en hun mogelijke consequenties roepen de vraag op aan welke methode de voorkeur zou moeten worden gegeven. In de jaren zeventig en de beginjaren tachtig bestond er onder gezondheidseconomen een vrij algemene consensus dat de standard gamble superieur was aan de overige methodes. De resultaten van ratioschaal en time trade-

off zouden beoordeeld moeten worden aan de hand van de uitkomsten van de standard gamble. De reden voor deze vermeende superioriteit van de standard gamble was dat de standard gamble gebaseerd is op een axiomatische theorie van beslissen: von Neumann en Morgenstern's verwachte nutstheorie. Het voordeel van een axiomatische theorie van beslissen is dat de aannames waaraan het keuzegedrag van een individu moet voldoen om zich volgens de theorie te gedragen, zijn geïdentificeerd. In het geval van het verwachte nutsmodel bestond de overtuiging dat de aannames dermate plausibel waren dat rationele individuen zich niet alleen volgens dit model zouden moeten gedragen, maar ook dat het model het keuzegedrag van individuen op een correcte wijze beschrijft.

De vermeende validiteit van het verwachte nutsmodel is echter gedurende de laatste decennia betwist. Empirische studies hebben aangetoond dat individuen zich in een aantal beslissingssituaties niet volgens het verwachte nutsmodel gedragen. Onderzoek van Lewellyn-Thomas et al. (1982) en van Rutten-van Mölken et al. (1995) heeft aangetoond dat het verwachte nutsmodel ook als verklaring van keuzes met betrekking tot gezondheidszorg tot inconsistente resultaten kan leiden. De empirische schendingen van het verwachte nutsmodel hebben ertoe geleid dat het vertrouwen in de standard gamble als gouden standaard van utiliteitsmeting is aangetast. Noch de ratioschaal noch de time trade-off heeft de standard gamble als norm kunnen vervangen. Dit heeft tot verwarring onder gezondheidsonderzoekers geleid: aan de ene kant is bekend dat de verschillende methodes tot verschillende resultaten leiden, aan de andere kant bestaat weinig inzicht in welke methode de voorkeur verdient.

Hoewel vergelijking (1) het centrale idee achter QALYs weergeeft (dat levensjaren gewogen moeten worden voor kwaliteit van leven), is het niet het model dat het meest gebruikt is in de literatuur. Gezondheidsprogramma's verschillen in het algemeen in het tijdstip waarop kosten en baten gerealiseerd worden. In het voorbeeld van de behandeling voor huidziekte leidt behandeling 1 sneller tot een verbeterde gezondheid dan behandeling 2. De meeste kosten-utiliteitsanalyses corrigeren uitkomsten voor het tijdstip waarop ze gerealiseerd worden. De reden hiervoor is dat het verloop van de tijd van invloed is op de aantrekkelijkheid van uitkomsten. Economen refereren aan dit fenomeen als tijdsvoorkeur. In het algemeen geldt dat naarmate een positieve uitkomst later in de tijd wordt gerealiseerd, deze uitkomst als minder aantrekkelijk wordt ervaren. De meeste mensen prefereren bijvoorbeeld het ontvangen van tien gulden nu boven het ontvangen van tien gulden over een jaar. Het verschijnsel dat (positieve) uitkomsten aantrekkelijker worden naarmate ze eerder gerealiseerd worden, wordt aangeduid met de term positieve tijdsvoorkeur. Negatieve tijdsvoorkeur staat dan voor het verschijnsel dat (positieve) uitkomsten aantrekkelijker naarmate ze later gerealiseerd

worden. Tijdsvoorkeur wordt meegenomen in kosten-utiliteitsanalyses door het toepassen van een constante disconteringsvoet *(r)*. De aanpassing voor tijdsvoorkeur leidt tot vervanging van vergelijking (1) door:

$$\sum_{t=1}^{T} \frac{u(q_t)}{(1+r)^{t-1}} \qquad (2)$$

Stel dat we in het bovenstaande voorbeeld besloten hebben dat een disconteringsvoet van 5% per jaar moet worden toegepast. Dat wil zeggen: r is gelijk aan 0.05. Behandeling 1 resulteert dan in 6.49 QALYs en behandeling 2 in 7.13 QALYs. Behandeling 2 wordt ook op basis van het criterium "maximaliseer het aantal 5% verdisconteerde QALYs" geprefereerd boven behandeling 1. Het verschil tussen de twee behandelingen is echter afgenomen, hetgeen te verklaren is door het feit dat behandeling 2 pas na twee jaar aantrekkelijker wordt dan behandeling 1.

Tot nog toe is gesproken over QALYs als beslismodel op het individuele niveau. Aan QALYs wordt ook een belangrijke rol toegedicht als het gaat om beleidsbeslissingen met betrekking tot de verdeling van middelen over gezondheidszorgprogramma's. In een vergelijking tussen gezondheidszorg-programma's zou dat programma gekozen moeten worden dat het grootste aantal QALYs voor een gegeven budget genereert. Gezondheidszorgprogramma's hebben in het algemeen betrekking op meerdere individuen. Om tot een zinvolle vergelijking van programma's te komen, moet daarom een procedure bepaald worden om de gezondheidsbaten van verschillende individuen te aggregeren. De wijze waarop dit in kosten-utiliteitsonderzoek gedaan wordt, is door ongewogen sommatie over alle relevante individuen. Wanneer het totale aantal individuen waarvan de gezondheid door de invoering van een gezondheidsprogramma beïnvloed wordt, gelijk is aan I, kan het aantal QALYs dat door een programma gegenereerd wordt berekend worden als:

$$\sum_{i=1}^{I} \sum_{t_i=1}^{T_i} \frac{u(q_{t_i})}{(1+r)^{t_i-1}} \qquad (3)$$

Het onderschrift i in t_i geeft weer dat het aantal levensjaren waarop het gezondheidsprogramma betrekking heeft niet noodzakelijkerwijs gelijk hoeft te zijn voor alle individuen. In principe kan ook u variëren over individuen (dat wil zeggen u zou vervangen moeten worden door u_i in vergelijking (3)). In praktische toepassingen wordt echter meestal de gemiddelde waarde van u gehanteerd en wordt impliciet verondersteld dat deze voor alle individuen geldig is.

Stel in het voorbeeld van de behandeling voor huidziekte dat behandeling 1 goedkoper is dan behandeling 2. Voor elke 1000 individuen die behandeling 2 ontvangen, kunnen 1075 individuen behandeling 1 ontvangen. Veronderstel dat alle individuen nog 10 jaar leven en abstraheer van eventueel noodzakelijke behandeling na 10 jaar. Veronderstel tenslotte dat de gemiddelde utiliteiten gelijk zijn aan 0.50 voor de ernstige vorm van huidziekte, aan 0.80 voor de lichte vorm van huidziekte en aan 1 voor geen huidziekte. Het disconteringspercentage tenslotte is gelijk aan 5%. Dan kan aan de hand van vergelijking (3) berekend worden dat voor een gegeven budget programma 1, waarin 1075 patiënten behandeling 1 ontvangen, 6976.75 QALYs oplevert. Programma 2, waarin 1000 patiënten behandeling 2 ontvangen, levert 7130 QALYs op. Het tweede programma wordt geprefereerd op basis van het criterium "maximaliseer het totale aantal verdisconteerde QALYs voor een gegeven budget".

2. Onderzoeksvragen

Sinds de introductie van kosten-utiliteitsanalyse halverwege de jaren zeventig is het aantal toepassingen van QALYs sterk toegenomen. Ondanks de gestegen populariteit van QALYs, bleef onduidelijkheid bestaan met betrekking tot de economisch theoretische grondslagen van QALYs als beslismodel. In dit proefschrift is getracht QALYs een fundering te geven binnen de besliskunde. Centraal in de besliskunde staan voorkeursrelaties. Om deze voorkeursrelaties hanteerbaar te maken, worden condities opgelegd. Deze condities maken het mogelijk om de voorkeursrelaties door middel van een model te beschrijven. In het voorbeeld van de huidziekte werden vergelijkingen (1) en (2) als model genomen om de keuzes van een individu met betrekking tot twee behandelingen te beschrijven. Deze twee vergelijkingen zullen de keuzes echter slechts dan correct voorspellen wanneer aan een aantal condities (axioma's) is voldaan. Een doelstelling van dit proefschrift is het identificeren van de condities die garanderen dat QALYs individuele voorkeuren correct weergeven. Het voordeel van het identificeren van condities die aan een voorkeursrelatie moeten worden opgelegd om tot een correct voorspellend model te komen, is dat dit de empirische beoordeling van het model mogelijk maakt. Het identificeren van de condities maakt een beoordeling van de normatieve en descriptieve validiteit van het model mogelijk. Normatieve validiteit heeft betrekking op de vraag of het redelijk is voor een individu om zich te gedragen volgens de condities. Descriptieve validiteit heeft betrekking op de vraag of de condities het gedrag van een individu correct beschrijft. Een voorbeeld kan het onderscheid tussen normatieve validiteit en descriptieve validiteit verduidelijken.

Stel dat we de condities geïdentificeerd hebben onder welke de keuzes van een individu beschreven kunnen worden aan de hand van vergelijking (1). Stel dat één van de geïdentificeerde condities als volgt luidt: als het individu 10 jaar zonder huidziekte prefereert boven 10 jaar met lichte huidziekte en hij prefereert 10 jaar met lichte huidziekte boven 10 jaar met ernstige huidziekte, dan moet hij ook 10 jaar zonder huidziekte prefereren boven 10 jaar met ernstige huidziekte. Deze conditie staat bekend onder de naam transitiviteit. Stel dat we na bestudering van de conditie besluiten dat het redelijk is voor een individu om zich volgens deze conditie te gedragen. In dat geval concluderen we dat transitiviteit normatief valide is. Bij gevolg is een model dat alleen transitiviteit oplegt aan de voorkeursrelatie ook normatief valide. Dit betekent niet dat een individu zich ook volgens transitiviteit zal gedragen. Het is goed mogelijk dat in beslissingssituaties dit individu transitiviteit systematisch schendt zelfs na herhaaldelijke uitleg dat zijn keuzes niet in overeenstemming zijn met transitiviteit. In dat geval concluderen we dat transitiviteit niet descriptief valide is: transitiviteit geeft geen goede beschrijving van de individuele voorkeuren in deze beslissingscontext.

Dit proefschrift behandelt QALYs als een nutsmodel. Hoofdstukken 2 tot en met 8 behandelen individuele voorkeursrelaties. Het is evident dat de resultaten van deze hoofdstukken relevant zijn voor besliskundige vraagstukken waarin het gaat om het verklaren en voorspellen van de keuzes van individuele patiënten. De relevantie van deze hoofdstukken voor beslissingen op het niveau van de samenleving vraagt enige toelichting. Een sociale voorkeursrelatie staat centraal in de laatste beslissingscontext en het is niet a priori duidelijk wat de relevantie van individuele voorkeuren is voor deze sociale voorkeursrelatie. Twee interpretaties van QALYs als sociale beslissingsregel kunnen grofweg onderscheiden worden in de literatuur: QALYs als maat van gezondheid en QALYs als maat van nut. Hoewel deze twee interpretaties elkaar niet noodzakelijkerwijs uitsluiten (QALYs als maat van nut kunnen ook een maat van gezondheid zijn), verschillen de interpretaties van QALYs in het belang dat aan individuele voorkeuren wordt toegekend. In de "QALYs als maat van gezondheid" interpretatie is er geen relatie tussen QALYs en individuele voorkeuren. In de "QALYs als maat van nut" interpretatie worden de utiliteiten die aan gezondheidstoestanden worden toegekend berekend aan de hand van individuele voorkeuren. Het is daarom evident dat een studie van individuele voorkeursrelaties relevant is voor de interpretatie van QALYs als maat van nut.

Hoewel de resultaten van dit proefschrift met name relevant zijn voor de interpretatie van QALYs als maat van nut, is de analyse van hoofdstuk 9, waarin de sociale voorkeursrelatie centraal staat, is ook toepasbaar in de interpretatie van QALYs als maat van gezondheid.

Gegeven de invalshoek van dit proefschrift, waarin QALYs als nutsmodel worden opgevat, staan vier vragen centraal:

1. Onder welke condities is het QALY model zoals weergegeven in vergelijking (1) een valide representatie van individueel keuzegedrag met betrekking tot gezondheid? Zijn deze condities normatief en descriptief valide? Zo niet, zijn alternatieve modellen meer valide?
2. Welke methode om utiliteiten voor gezondheidstoestanden te bepalen (de ratioschaal, de time trade-off en de standard gamble) genereert resultaten die individuele keuzes het meest accuraat beschrijven?
3. Welke condities moeten worden opgelegd om de individuele intertemporele voorkeursrelatie te kunnen weergeven door middel van vergelijking (2)? Zijn deze condities normatief en descriptief valide? Zo niet, zijn alternatieve modellen meer valide?
4. Welke condities moeten aan de sociale voorkeursrelatie worden opgelegd om deze te kunnen weergeven door middel van vergelijking (3)? Zijn deze condities een correcte reflectie van sociale voorkeuren? Zo niet, zijn er alternatieve sociale beslissingsregels die de sociale voorkeuren beter weergeven?

3. Resultaten

3.1 Validiteit van QALYs

De eerste van de hierboven geformuleerde vragen is bestudeerd in hoofdstukken 2 tot en met 4. Hoofdstukken 2 en 3 hebben een theoretisch karakter en behandelen de vraag welke condities aan individuele voorkeuren moeten worden opgelegd om individueel keuzegedrag door middel van QALYs te kunnen representeren. Hoofdstuk 2 behandelt de situatie waarin alle gezondheidstoestanden chronisch zijn, dat wil zeggen de gezondheid van een individu is constant over de tijd. Een voorbeeld van een chronische toestand is behandeling 1 in de keuze van behandeling voor huidziekte. Behandeling 1 resulteert in 10 jaar met een lichte vorm van huidziekte. Pliskin et al. (1980) hebben in een eerder artikel reeds afgeleid onder welke condities het QALY model individuele voorkeuren met betrekking tot chronische gezondheidstoestanden correct weergeeft. In hoofdstuk 2 is aangetoond dat twee van de drie condities die Pliskin et al. opleggen niet noodzakelijk zijn. Naast een zwakke conditie, die zeer plausibel is in de medische context, is *risiko neutraliteit met betrekking tot levensjaren* de enige van de drie door Pliskin et al. geïdentificeerde condities die hoeft te worden opgelegd. Risiko neutraliteit met

betrekking tot levensjaren is een conditie die zowel eenvoudig te begrijpen als empirisch makkelijk te testen is. Dit vereenvoudigt de normatieve en descriptieve beoordeling van het model. Het feit dat QALYs aan de hand van een transparante conditie gekarakteriseerd kunnen worden is de kracht van het in hoofdstuk 2 afgeleide resultaat. Dit betekent overigens niet dat hoofdstuk 2 een rechtvaardiging bevat voor het gebruik van QALYs. Risiko neutraliteit met betrekking tot levensjaren is een restrictieve conditie. Empirisch onderzoek heeft in het algemeen aangetoond dat respondenten zich niet volgens deze conditie gedragen. Omdat risiko neutraliteit met betrekking tot levensjaren een eenvoudig te testen conditie is, verdient het aanbeveling om in kosten-utiliteitsanalyses te toetsen of respondenten zich volgens deze conditie gedragen. Wanneer respondenten risiko neutraliteit met betrekking tot levensjaren systematisch schenden, verdient het aanbeveling alternatieve nutsmodellen, die meer algemeen zijn, te gebruiken. Een voorbeeld van een alternatief model is het zogenaamde "risk-adjusted QALY model" voorgesteld in Pliskin et al. In dit model worden levensjaren aangepast voor risikohouding. De twee condities waarop dit model is gebaseerd, zijn getest in hoofdstuk 4. De resultaten van hoofdstuk 4 zijn gunstig voor de descriptieve validiteit van het algemene model van Pliskin et al. Slechts een relatief laag percentage van de respondenten gedraagt zich exact volgens de condities. Hierbij moet bedacht worden dat respondenten in het algemeen onbekend zijn met de gezondheidstoestanden die zij gevraagd worden te waarderen en met de methodes om utiliteiten aan de gezondheidstoestanden toe te kennen. Dit leidt tot een zekere mate van imprecisie in de gegeven antwoorden. Na aanpassing voor deze imprecisie blijkt een meerderheid van de respondenten zich volgens de condities van het meer algemene model te gedragen.

In hoofdstuk 3 is de analyse van hoofdstuk 2 uitgebreid. Hoofdstuk 3 bevat een karakterisering van het QALY model voor de context waarin kwaliteit van leven kan variëren over de tijd. De centrale conditie die QALYs in deze context karakteriseert is vervolgens geëvalueerd aan de hand van uit de empirische literatuur bekende resultaten. De algemene conclusie die uit hoofdstuk 3 naar voren komt, is dat deze conditie restrictief is en in veel beslissingssituaties niet representatief is voor het keuzegedrag van individuen. Deze voor QALYs negatieve conclusie suggereert dat meer algemene modellen gebruikt moeten worden om individueel keuzegedrag met betrekking tot in kwaliteit variërende gezondheidsprofielen te verklaren. In hoofdstuk 3 is een meer algemeen model voorgesteld. Dit model generaliseert de QALY door de lineariteit met betrekking tot levensjaren op te geven.

Een andere generalisatie van de QALY is de door Mehrez en Gafni (1989) voorgestelde *"healthy-years equivalents"* (HYEs). De HYE is ontwikkeld in reactie op uit de empirische literatuur bekende schendingen van de condities die aan het

QALY model voor chronische gezondheidstoestanden ten grondslag liggen. Zoals boven al is opgemerkt, zwakken de resultaten van dit proefschrift uitspraken over vermeende schendingen van de QALY condities enigszins af. Na aanpassing voor imprecisie in voorkeuren voldoet een meerderheid van de respondenten aan de condities die aan het chronische QALY model ten grondslag liggen. Mehrez en Gafni hebben in meerdere publikaties beargumenteerd dat de HYE superieur is aan de QALY. Hun argument wordt gemotiveerd door de stelling dat de HYE geen condities oplegt aan de nutsfunctie van een individu. Hierdoor zou de HYE individuele voorkeuren altijd correct representeren. Een recent debat in de literatuur, waaraan hoofdstuk 3 een bijdrage is, heeft aangetoond dat deze stelling niet houdbaar is. De in brede kring aanvaarde uitkomst van dit debat is dat het gebruik van HYEs als uitkomstmaat in kosten-utiliteitsanalyse geen wezenlijke verbetering is ten opzichte van QALYs.

3.2 Methodes

In hoofdstukken 5 en 6 is de vraag behandeld aan welke methode van utiliteitsmeting de voorkeur gegeven dient te worden wanneer het doel is individuele voorkeuren zo accuraat mogelijk te beschrijven. In hoofdstuk 5 staat een waargenomen inconsistentie in standard gamble antwoorden centraal: standard gambles die volgens het verwachte nutsmodel tot identieke utiliteiten zouden moeten leiden, leiden tot verschillende utiliteiten. Deze inconsistentie is eerder waargenomen door Llewellyn-Thomas et al. (1982). In hoofdstuk 5 is getracht een verklaring voor deze inconsistentie te geven. Drie verklaringen zijn onderzocht. De eerste verklaring is dat de inconsistentie een gevolg is van het feit dat de vorm van de standard gamble verschilt. De tweede verklaring is dat de inconsistentie veroorzaakt wordt door imprecieze voorkeuren van respondenten. De derde verklaring is dat respondenten kansen wegen en niet lineair evalueren zoals het verwachte nutsmodel impliceert. Uit de analyse van hoofdstuk 5 blijkt dat kansweging de voornaamste verklaring van de inconsistentie is. De respondenten die aan het in hoofdstuk gerapporteerde experiment deelnamen, blijken zeer pessimistisch te zijn in de zin dat aan de kans op een succesvolle uitkomst van behandeling een laag gewicht wordt toegekend en aan de kans op een mislukking van behandeling een hoog gewicht. Dat wil zeggen: respondenten zijn zeer afkerig van het lopen van risiko. De geschatte kanswegingsfunctie wijkt significant af van de door het verwachte nutsmodel voorspelde lineariteit.

Een verrassend resultaat van de in hoofdstuk 5 gepresenteerde analyse is dat de geschatte kanswegingsfunctie sterk afwijkt van met betrekking tot monetaire

uitkomsten geschatte kanswegingsfuncties. De door Tversky en Kahneman (1992) geschatte kanswegingsfunctie blijkt de resultaten van het in hoofdstuk 5 beschreven experiment zelfs minder goed te verklaren dan de lineaire functie die in het verwachte nutsmodel wordt gebruikt. De slechte verklaring die de kanswegingsfunctie van Tversky en Kahneman geeft, kan in de eerste plaats het gevolg zijn van het speciale karakter van gezondheid als uitkomst. De afwijking kan echter ook het gevolg zijn van de wijze waarop de standard gamble in het algemeen gesteld wordt in gezondheidsonderzoek. In gezondheidsonderzoek wordt aan respondenten gevraagd voor welke kans op succesvolle behandeling zij indifferent zijn tussen twee loterijen. Uit de empirische literatuur is bekend dat een standard gamble waarin gevraagd wordt indifferentie in termen van kansen uit te drukken tot te hoge utiliteiten leidt [zie bijvoorbeeld Hershey en Schoemaker (1985)]. De sterk pessimistische houding die geobserveerd wordt in hoofdstuk 5 kan ook een gevolg zijn van een systematische overschatting van de utiliteiten, die voortvloeit uit de wijze waarop de standard gamble vragen werden gesteld.

Een implicatie van de aanwezigheid van kansweging is dat standard gamble antwoorden moeten worden aangepast voor kansweging. Het achterwege laten van deze aanpassing geeft misleidende resultaten. Bij mijn weten is hoofdstuk 5 de eerste poging om tot een schatting van een kanswegingsfunctie met betrekking tot gezondheid te komen. Uit het bovenstaande mag duidelijk zijn dat verder onderzoek naar de precieze vorm van de kanswegingsfunctie voor gezondheid noodzakelijk is.

De in hoofdstuk 6 gepresenteerde analyse bevestigt de conclusie dat de standard gamble tot te hoge utiliteiten leidt. Gegeven de centrale doelstelling van dit proefschrift om keuzes te verklaren, zijn in hoofdstuk 6 QALYs berekend aan de hand van ratioschaal gewichten, time trade-off gewichten en standard gamble gewichten vergeleken met feitelijke keuzes. De voornaamste conclusies van hoofdstuk 6 zijn dat de standard gamble tot te hoge gewichten voor gezondheidstoestanden leidt, dat de ratio-schaal tot te lage gewichten leidt en dat QALYs berekend aan de hand van time trade-off gewichten (TTO-QALYs) feitelijke keuzes het meest accuraat weergeven. De in hoofdstuk 6 gepresenteerde analyse suggereert dat de time trade-off gebruikt zou moeten worden bij de bepaling van utiliteiten voor gezondheidstoestanden. Er moet echter een kwalificatie bij deze conclusie worden gemaakt. Uit de in hoofdstuk 6 gepresenteerde analyse blijkt tevens dat wanneer discontering wordt toegepast, TTO-QALYs individuele keuzes minder accuraat beschrijven. Dit patroon is consistent met een hypothese geformuleerd door Stalmeier et al. (1995) dat individuen een bepaalde vuistregel hanteren bij de beantwoording van time trade-off vragen. Bedacht moet worden dat een vuistregel die in het algemeen goed werkt niet in elke beslissingssituatie

voorkeuren correct weergeeft. Een vuistregel is niet gelijk aan werkelijke voorkeuren en kan tot misleidende resultaten leiden. Niettegenstaande deze kwalificatie ondersteunen de resultaten van dit proefschrift het gebruik van de time trade-off bij de bepaling van utiliteiten voor gezondheidstoestanden.

3.3 Tijdsvoorkeur

Hoofdstukken 7 en 8 behandelen intertemporele voorkeuren voor gezondheid. Centraal staat de vraag naar de meest geschikte wijze om intertemporele voorkeuren voor gezondheid te modelleren. Hoofdstuk 7 bevat een theoretische verhandeling over het in kosten-utiliteitsanalyse meest gebruikte model waarin tegen een constant percentage verdisconteerd wordt. De condities die aan dit model ten grondslag liggen, zijn geïdentificeerd. Argumenten zijn aangevoerd waarom zowel de normatieve als de descriptieve validiteit van deze condities omstreden is. Een generalisatie van een model met een constant disconteringspercentage is een model waarin het disconteringspercentage kan variëren over de tijd. De karakterisering van dit model volgt eenvoudig uit de karakterisering van het model met een constant disconteringspercentage. Het is evident dat een model met een variabel disconteringspercentage beter individuele voorkeuren verklaart dan een model met een constant disconteringspercentage. Dit volgt uit het feit dat een model met een variabel disconteringspercentage minder restricties oplegt aan de voorkeursrelatie. In hoofdstuk 7 is echter beargumenteerd dat modellen met een variabel disconteringspercentage een geheel eigen probleem kennen: individuen die zich volgens een variabel disconteringsmodel gedragen, kunnen inconsistent zijn in de zin dat hun keuze over de tijd verandert. Zulk gedrag wordt omschreven met de term "dynamische inconsistentie". De mogelijkheid van dynamisch inconsistent gedrag maakt de normatieve validiteit van intertemporele modellen met een variabel disconteringspercentage omstreden.

Hoofdstuk 8 bevat een empirische test van de centrale conditie van modellen waarin een constant disconteringspercentage gehanteerd wordt. Aan de hand van de resultaten van hoofdstuk 8 moet deze conditie verworpen worden. Uit de analyse van hoofdstuk 8 blijkt dat individuen zich ook met betrekking tot gezondheidsuitkomsten gedragen volgens een principe dat Loewenstein en Prelec (1992) omschrijven als het "common difference effect". Loewenstein en Prelec identificeerden dit principe met betrekking tot monetaire uitkomsten. In hoofdstuk 8 is aangetoond dat hetzelfde principe keuzes met betrekking tot gezondheid bepaalt.

Aan de hand van het "common difference effect" kunnen alternatieve modellen van intertemporeel keuzegedrag geformuleerd worden. Deze gegeneraliseerde disconteringsmodellen zijn algemener dan modellen met een constant disconteringspercentage. Ook in deze modellen bestaat echter de mogelijkheid van dynamisch inconsistent gedrag. Bovendien veronderstellen deze gegeneraliseerde disconteringsmodellen, net als modellen met een constant disconteringspercentage, separabiliteit van voorkeuren over de tijd. Dat wil zeggen dat de aantrekkelijkheid van het verkrijgen van een uitkomst nu niet beïnvloed wordt door wat in het verleden heeft plaatsgevonden noch door wat in de toekomst zal plaatsvinden. Het mag duidelijk zijn dat dit een stringente veronderstelling is, die niet in elke beslissingscontext geldig is. Hoofdstuk 7 bevat verschillende suggesties hoe intertemporele afhankelijkheid gemodelleerd zou kunnen worden. Een evident nadeel van deze modellen is dat de beschrijving van individuele voorkeuren gecompliceerder wordt. Er zal een goede afweging tussen nauwkeurigheid van beschrijving en praktische hanteerbaarheid gevonden moeten worden. Dit is een belangrijk gebied waarop toekomstig onderzoek zich kan richten.

3.4 Sociale voorkeuren

Hoofdstuk 9 behandelt de rol van QALYs in beleidsbeslissingen. Centraal staat de sociale voorkeursrelatie met betrekking tot de verdeling van QALYs (of andere uitkomstmaten) over individuen. Zoals boven al is uiteengezet, is ongewogen sommatie van individuele QALYs de gebruikelijke procedure in kosten-utiliteitsanalyse om het totaal aantal QALYs van een gezondheidszorgprogramma te bepalen. De vraag is of deze procedure in overeenstemming is met sociale voorkeuren ten aanzien van verdelingsvraagstukken. Voordat deze laatste vraag op zinvolle wijze beantwoord kan worden, is een rechtvaardiging gegeven waarom QALYs over individuen geaggregeerd mogen worden. Met name in het geval van de QALYs als maat van nut interpretatie is een rechtvaardiging noodzakelijk. De aggregatie van nut over individuen is een zeer omstreden onderwerp binnen de economie.

De in hoofdstuk 9 gepresenteerde analyse laat zien dat ongewogen sommatie bepaalde rechtvaardigheidsoverwegingen uitsluit. Ongewogen sommatie laat bijvoorbeeld geen ruimte voor voorkeuren met betrekking tot de verdeling van QALYs. Het maakt niet uit of één individu 100 QALYs ontvangt of dat honderd individuen ieder 1 QALY ontvangen. In bepaalde beslissingssituaties zullen verdelingsvraagstukken een rol kunnen spelen. In dergelijke situaties ontstaat de behoefte aan beslissingsregels die rechtvaardigheidsoverwegingen kunnen

meenemen. Tot op heden zijn in de gezondheidseconomie nauwelijks pogingen ondernomen om dergelijke beslissingsregels te ontwikkelen. Hoofdstuk 9 bevat een aantal beslissingsregels die het mogelijk maken rechtvaardigheidsprincipes in de besluitvorming op te nemen. Een belangrijke achterliggende overweging bij de ontwikkeling van deze alternatieve beslissingsregels is praktische toepasbaarheid. De regels mochten niet dermate gecompliceerd zijn dat praktische toepasbaarheid vrijwel onmogelijk werd gemaakt. Een gevolg van dit uitgangspunt is dat de beslisregels gebaseerd zijn op condities die niet in elke beslissingscontext realistisch zijn. Het verder afzwakken van deze condities zal tot een betere beschrijving van sociale voorkeuren leiden, maar zal ten koste gaan van de praktische toepasbaarheid.

4. Tot slot

Doel van dit proefschrift is het leveren van een bijdrage aan de oplossing van methodologische problemen in kosten-utiliteitsanalyse. Zonder te pretenderen compleet te zijn geweest, geloof ik dat dit proefschrift bijdraagt aan het inzicht in de theoretische basis van op nutstheorie gebaseerde uitkomstmaten in gezondheidsonderzoek. De theoretische resultaten die in dit proefschrift zijn gepresenteerd helpen bij de normatieve en descriptieve beoordeling van het gebruik van utiliteitsmaten in het algemeen en QALYs in het bijzonder en geven inzicht in welke beslissingssituaties deze utiliteitsmaten bruikbaar zijn. De empirische studies die in dit proefschrift zijn beschreven bieden verder inzicht in de descriptieve validiteit van de verschillende modellen.

De algemene boodschap van dit proefschrift is gematigd positief voor op QALYs gebaseerde besluitvorming. Hoewel in de theoretische analyses regelmatig twijfel wordt uitgesproken over de validiteit van de aannames, blijkt in het empirische gedeelte dat voor een aantal van de aannames een redelijke mate van steun bestaat. Meer onderzoek moet worden uitgevoerd naar de toepasbaarheid van utiliteitsmaten in gezondheidszorgonderzoek. De resultaten van dit proefschrift zijn echter bemoedigend voor de richting die gezondheidszorgonderzoek in de afgelopen decennia is ingeslagen.

Acknowledgements

Several persons have provided helpful comments on the papers that constitute this thesis. In particular I should like to thank my two supervisors Eddy van Doorslaer and Peter Wakker. Eddy van Doorslaer initiated the research that led up to this thesis. Even though I fear that at times I overestimated his ability to grasp within a few "minutes" ideas I had been working on for months, Eddy always managed to provide useful comments. It probably helped that most of our discussions took place at locations where a glass of beer was within reach. Peter Wakker has without any doubt most strongly influenced my thinking on the topics reported in this thesis. I much appreciate his patience in explaining me the intricacies of utility theory.

In spite of their crammed time schedules Graham Loomes, Amiram Gafni, Magnus Johannesson and Christophe Gonzales were always willing to read through various drafts of papers and to pinpoint inadequacies. Thanks.

Thanks to Jaco van Rijn for his assistance in running several of the experimental sessions, the results of which are reported in this thesis and to Maureen Rutten-van Mölken for help in the selection of health states. Machiel Crielaard deserves credit for his design of the cover of this thesis.

I am grateful to Merck, Sharpe and Dome for financial support.

Finally a word of thanks to family, friends and colleagues who kept me aware of the fact that at times it is very appealing to reveal contradictory trade-offs.