Paulus Franciscus Maria Krabbe

The valuation of health outcomes: a contribution to the QALY approach
Thesis Erasmus University Rotterdam
ISBN 90-9011529-3
NUGI 743
key words: health status, QALY model, valuation methods
date of issue: June, 1998

Address for correspondence:
P.F.M. Krabbe
Department of Medical Informatics, Epidemiology & Statistics
Medical Technology Assessment unit (152 MIES)
Faculty of Medical Sciences
University of Nijmegen
P.O. Box 9101
6500 HB Nijmegen
The Netherlands

Cover:
representation of the firmament created by a mathematical algorithm (printed in black and one separate bleu color)

THE VALUATION OF HEALTH OUTCOMES

A Contribution to the QALY Approach

DE WAARDERING VAN GEZONDHEIDSUITKOMSTEN

Een bijdrage tot de QALY benadering

Proefschrift

TER VERKIJGING VAN DE GRAAD VAN DOCTOR
AAN DE ERASMUS UNIVERSITEIT ROTTERDAM
OP GEZAG VAN 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
WOENSDAG 2 SEPTEMBER 1998 OM 13.45 UUR.

door

Paulus Franciscus Maria Krabbe

geboren te Heemskerk

Promotiecommissie

Promotor
Prof. dr J.D.F. Habbema

Overige leden
Prof. dr J. Kievit
Prof. dr J. Passchier
Prof. dr F.F.H. Rutten

Co-promotor
dr G.J. Bonsel

This thesis was financially supported by

Department of Public Health (Instituut Maatschappelijke Gezondheidszorg), Erasmus University Rotterdam

Dutch Cancer Society (Nederlandse Kankerbestrijding – Koningin Wilhelmina Fonds)

Johan Vermeij Foundation (Johan Vermeij Stichting)

Theta Research

Indien er enige gave van de goden is aan de mensen,
is het redelijk aan te nemen, dat het geluk een geschenk
der goden is en wel vóór alle andere menselijke goederen,
inzover dit het beste is van alle.
Maar misschien behoort deze kwestie meer thuis in een
ander onderzoek; dit evenwel is duidelijk, dat het geluk,
ook al is het niet door de goden ons gezonden, maar verkregen
door deugdzaamheid, studie of oefening,
toch tot de meest goddelijke dingen behoort.

ARISTOTELES
(384 – 322 B.C.)

Contents

1. Introduction 9
2. On the quantification of health 13

PART I — SOCIETAL PERSPECTIVE

3. On the equivalence of collectively and individually collected responses 51
4. The comparability and reliability of five health-state valuation methods 71
5. Sequence effects, health profiles and the QALY model 91
6. The effect of adding a cognitive dimension to the EuroQol multiattribute health-status classification system 105

PART II — INDIVIDUAL PERSPECTIVE

7. The impact of clinical information on patients' preferences for cancer treatment 121
8. Analytical versus holistic approaches to the appraisal of treatments 135
9. The stability of outcome preferences 149

10. Discussion and conclusions 163

Appendix A 183
Appendix B 195
Appendix C 197
References 203
Summary 217
Samenvatting 223
Co-authors 229
List of scientific publications 230
Glossary 231
Dankwoord 237
Curriculum Vitae 239

Chapter 1

Introduction

The Subject Matter

The aim of any medical intervention is to improve or sustain health. Therefore, the assessment of health outcome is an important aspect in the scientific evaluation of medical effectiveness. Nowadays this is an important activity. Three major aspects of health outcome can be distinguished. The most prominent is survival, which defines the prevention of premature death as a major goal of care. Disease-specific clinical characteristics of the seriousness of the pathology are a second aspect of health outcome. More recently, a third aspect has been developed to express an entity that was previously more or less implicit: health status. Especially in the more prosperous part of the world where life expectancy has shown an impressive increase, a shift of emphasis from the attention of quantity of life towards the quality of health, i.e., to health status can be noticed.

Health status can be thought of as complementary to survival in the evaluation of the impact of disease and of the effect of medical interventions. It should be noted that health status is not equivalent to the broader concept of "quality-of-life" or "well-being", though it may be seen as synonymous to the qualified concept "health-related quality-of-life". Quality-of-life depends upon factors other than health, such as economic status and education. Its appraisal may be affected by such aspects as spirituality, culture, personality and political systems. Quality-of-life obviously can be strongly determined by health status, nevertheless a good health status forms no guarantee of being »glücklich«.

The operationalization and subsequent empirical assessment of health status consists of several steps. First, a decision has to be made on the comprehensiveness, i.e., the contents of health status. Health status is generally considered to comprise three main domains in relation to health: physical, psychological and social functioning. The second step consists of a description of the health status of individuals or groups. Its measurement is usually operationalized by the use of health-status questionnaires (Essink-Bot, 1995).*

This thesis, however, is explicitly focused on the third step, i.e., the *valuation* of health outcome. As part of clinical decision analysis and cost-utility analysis, the valuation of health outcome has gained wide attention. Essentially, after the operationalization of a health concept, the procedure involves the explicit valuation of one or more health outcomes by a person or panel. The usual way of presenting health outcomes to the subjects who perform the valuation task is by means of shorthand descriptions of representative health outcomes. More precisely, such scenarios consist of one or more statements on health status with or without reference to duration. Subsequently, health status and survival are conventionally merged into a single summary value, the number of so-called Quality-Adjusted Life Years (QALYs).

---

* We will refer to the *valuation* of health outcome if the assessment task is solely focused on the assignment of a single number (value) to a particular hypothetical health outcome. Descriptive health-status *measures* are conceptually different and deal with description of own health status, e.g., by means of self-assessment questionnaires.

During the last twenty-five years, considerable effort has been invested in the development of procedures for the valuation of health outcomes. Such valuations are vital to be able to deal with various research questions. They are essential components of the information used in rational decision making in health care, both at the level of individual treatment choices and at the public health policy level. Another goal for which accurate valuations for a broad range of different health outcomes are needed, is the monitoring of the health of the population. Scientists, physicians and policy makers alike, are eager to have access to methods which yield reliable valuations for specific health outcomes.

Validity and reliability are two basic requirements for empirical valuations of health outcomes. When using the term *validity* we refer to the degree to which a method really measures what it intends to measure. Especially for (partly) subjective phenomena, such as health, the determination of validity is not easy. As opposed to phenomena such as temperature, blood pressure or survival, health status is not directly observable and its appraisal is basically idiosyncratic (normative). For this reason, specific measurement procedures are required. The second requirement, *reliability*, deals with the stability or repeatability of measurements. It is hoped that this thesis contributes to the study of these two psychometric aspects within the context of health-outcome valuation.

Validity and reliability are treated as two distinct methodological concepts although mathematically and conceptually the boundary is blurred. Achieving reliability is to some extent a technical matter. Validity, however, is much more than applying the appropriate technique. It is closely related to the essence of science itself, since it is concerned with the nature of "reality" and the nature of the properties measured (Kerlinger, 1986).

We consider that both conceptually and empirically, valid measurement procedures for health outcome are possible and feasible, such as in other domains of the social sciences which deal with entities which cannot be measured directly. This thesis is primarily directed at the general question of how to validly and reliably quantify health and health outcomes. The general QALY model is at the center of the answer to this question.

Most concepts and techniques discussed in this thesis are not only applicable to medicine. They may be found also in policy research for decision making in non-medical issues like expanding airports, location of oil drilling facilities and other decision problems that deal with multiple attributes requiring valuation.

It is the elusiveness of life, disease and death which renders the subject matter its enduring attractiveness and importance for health scientists.

Objectives and Structure

The focus of this thesis is on methodological issues in the valuation of health outcomes. The main objectives are:

- To provide an overview of the methodological state of the art in the field of valuation of health outcomes.
- To describe some variants of the general QALY mode altogether with their assumptions.
- To determine the validity and reliability of values for health outcomes and health profiles elicited by several different valuation methods.
- To provide insight into factors which may affect the valuation of health outcomes.

The thesis is based on the results of three groups of empirical studies. Chapter 3 to Chapter 5 are based on a research project entitled "Standardization in Medical Technology Assessment"; Chapter 6 is based on the research project "Disability Weights for Diseases in The Netherlands"; Chapters 7 to 9 are based on the research project "Treatment Choice in Laryngeal Cancer".

The outline of the thesis is as follows. Chapter 2 is an introduction to the methodological issues and conceptual models related to the valuation of health outcomes. In Chapter 3 aggregated individual valuations for health outcomes are compared with group values elicited by a collective voting system. Additionally, we examine the validity and reliability of the two valuation methods employed in this experiment. Chapter 4 deals in detail with the validity and reliability of five valuation methods. In Chapter 5, a key assumption of the conventional QALY model, additive utility independence, is examined. Validity aspects of the use of health-status scenarios, based on a multiattribute classification system, are studied in Chapter 6.

The next three chapters deal with individual assessments of health outcomes related to laryngeal cancer. Chapter 7 focuses on the effect of clinical information on the individual valuation of health outcomes. In Chapter 8, two different QALY models are compared. The issue of the stability of individual valuations is discussed in Chapter 9.

At the end, in Chapter 10, the main research problems related to the general QALY model are discussed together with their implications for the present state of the art in health outcome assessment.

Throughout this thesis we have aimed for a consistent use of (technical) terms. As the material of this thesis has evolved over time, this aim could not be achieved perfectly for some of the published articles. A summary and a glossary can be found at the end of this thesis.

Chapter 2

On the Quantification of Health

The Context of Health-Outcome Valuation

During the last twenty-five years there has been an increasing interest in the valuation of *health outcomes*. The aim of empirical valuation procedures is to capture a single numerical figure that represents the value of a particular health outcome. "Value" refers to the level of, say, desirability that a person associates with a particular health outcome, that is not necessarily his/her own. Implementation of such values ranges from the use of an individual patient's preference for treatment choices in clinical medicine, to the use of aggregate values in planning and monitoring of health programs. Other words that are frequently used as a synonym for value are: worth, weight, rating, function, utility, preference, preference value or score.

The above mentioned values are obtained by constructing, for each health outcome of interest separately, one (or more) scenarios in the form of characterizations or descriptions. After some editing to get an appropriate stimulus, the health-outcome scenario is subsequently valued. If the health-outcome scenario solely comprises a static health-status element, we refer to this as a *health-state* scenario. Often the resulting health-state scenario values are combined with values assigned to the *duration* of the health states to compute a comprehensive outcome measure for health as a whole. For the combination of those two elements of health, several so-called Quality-Adjusted Life Year (QALY) models have been developed.

The valuation of health outcomes, and of health status in particular, is a complex procedure, because health is part of life itself and its valuation depends on societal and personal value systems. Research on this topic is also complicated because it addresses concepts, operationalizations and applications which are used by different disciplines such as health economics, public health medicine, clinical epidemiology, psychology, and decision sciences. In addition, the interests of these disciplines are not similar. The emphasis of health economics is on unified, cost-related measures for the medical benefits of interventions. Public health is interested in an epidemiologically valid description of the health of a population and in measures of potential or actual change. Psychologists and sociologists emphasize the numerous distinctive factors that affect the process and results of individual evaluation.

This chapter on methodological aspects related to the valuation of health outcomes starts with a section on the different purposes for which these values are utilized. Following convention, we will refer to these values as health-state values rather than health-outcome values, although we are aware that this term does not completely cover the concept.

Purposes

Basically, we can distinguish two levels of specific applications of health-state values; according to their level of application, the individual level and the societal level respectively.

Individual level

At the individual level, studies are directed at treatment decisions in regard to an individual patient. Treatment here comprises any medical act like diagnostics, therapeutics, screening or just the provision of prognostic information. As a general approach, the patient's preference for a particular treatment option, taking into account the various possible outcomes, can be measured using hypothetical decision tasks which can subsequently be used as part of an analysis, formally referred to as as the discipline of clinical decision analysis.*

For the individual patient, a formal approach is superfluous if one treatment option is dominant, i.e., if for all relevant aspects one treatment option is clearly the best. Although absolute dominance on all relevant aspects will be rare in clinical practice, one treatment alternative usually dominates with respect to most aspects. If, however, this is not the case individual tradeoffs† usually concern aspects other than survival. Consider, for example, the choice of a drug when all drugs on choice have significant, yet different side-effects. Sometimes, tradeoffs between health status and survival are required, usually at some stage during the treatment of chronic progressive diseases, when only risky and disabling therapies may prolong life.

In this thesis, we discuss the individual tradeoff between speech problems and survival in the context of treatment decisions in laryngeal cancer. A scientific approach which is able to deal with individual tradeoffs is consistent with the call for more patient participation and reflects the attitude of providers that there should be more patient involvement in decision making.

Societal level

At the societal level, two major applications may be distinguished. First, the measurement of the impact of diseases on the health of populations; here we refer to the report of the World Bank (1993) and the report "Future scenario's of Public Health" (Ruwaard & Kramers, 1997), the latter carried out in the Netherlands. In this descriptive context of "burden of disease", several models have been developed to incorporate values for health states. No intervention or alternative is at issue but an artificial reference scenario, say, a general population without disease that invariably dies instantaneously at the age of 85. In recent years considerable progress has been made in developing a single public health measure that incorporates both mortality (life years) and morbidity (health status) of the population. By comparing the actual health of the population with the "reference health", population models may contribute to health planning and prioritizing.

Second, health-state values are important if specific health care decisions are an issue at regional or national level. The interest is not focused on the individual patient or on specific groups of patients but on the comparison and effectiveness of available health care interventions across diseases and across health care sectors. Decisions have to be made for example in regard to appropriate procedures and the use of existing and new technologies for particular treatments, specialties and diseases (Bonsel, 1991). Such health evaluation

---

* We are aware that this process in clinical practice usually has a more informal nature.

† A tradeoff is here defined as an informed weighting between non-dominant alternatives and a subsequent rational decision.

programs are part of Medical Technology Assessment (MTA) research. If efficiency is the guiding principle for choice, Cost-Effectiveness Analysis (CEA) and Cost-Utility Analysis (CUA) are the preferred strategies (Gold *et al.*, 1996; Drummond *et al.*, 1997). Both strategies incorporate the costs of medical interventions in their computations and express the added value of an intervention on health given a particular extra investment of resources. One basic aspect of CUA is the incorporation of values (utilities) for the health outcomes. The insurance coverage for assumed expensive new health care programs like organ transplantation and the pricing and licensing of new drugs on the market can be subjected to MTA studies, including CUA at a national level (see, e.g.,; Bonsel *et al.*, 1990; Elsinga & Rutten, 1997).

The use of health outcome values does not imply a fixed normative choice by science. Moreover, these values are used to support decision making in health care and to guide health planning in a systematic and standardized manner. At each level, the subject for which decisions have to be made relates to different aspects. Patient's decision making is involved with different optional treatments and their related prognosis, whereas studies at the societal level generally also incorporate the costs of the medical interventions in the analysis. Studies at the societal level may also consider distribution issues (social justice, equity considerations).

Applications that are focused on similar decisions as mentioned at the individual level, but for a patient group, are widely available. These applications assume relevant alternative health care strategies for such a group, and a decision has to be made regarding which strategy to pursue and which to defer or to reject. It is clear that an accurate group estimate of health outcome is easier to assess than an individual estimate. These types of studies will not be discussed separately, as they are closely related to studies at the societal level.

Summary values of health

Approaches to the individual and societal level of application have in common that they require a numerical expression of health outcome. Survival and health-status aspects have to be incorporated in a consistent way, in order to arrive at a summary value for health. A central assumption of this thesis is that a formalized, rational quantitative approach towards such health outcome values is possible and valid.

Conventionally, the description of any health outcome is based on two elements: "health status" (or "quality-of-health",i.e., "health-related quality-of-life") and "survival". These elements are integrated into a single measure, i.e. QALYs. Formally, we speak about QALY models. The operationalization of QALY models is complex (see: Appendix A). Quality-of-life or health status is formally denoted here as "Q" and survival (in most cases using the number of life years as the unit of measurement) is denoted as "L". In this thesis, we will address the various methodological problems associated with (this class of) health outcome valuations, which is usually expressed in units of measurement known as QALYs. In the next section the various stages in the process of arriving at QALYs will be presented.

A Five-Stage Approach Towards a Summary Value for Health Outcome

The process of obtaining a summary value for a health outcome, can be characterized as a five-stage approach. These five stages that each require a choice among several options, are the following.

I The QALY Model Stage.
II The Health Concept Stage.
III The Valuation Stage.
IV The Description Stage.
V The Analytical Stage.

These five stages should not be regarded as conceptually independent. A choice at one particular stage has consequences for the choices to be made at other stages (Essink-Bot, 1995), so the options should be considered carefully at the outset of the study. In this paragraph, we will briefly discuss the five stages of "QALY modeling", while in the next paragraphs, stages I, II, III and V will be discussed in more detail. Some relationships between the stages will be explained in subsequent paragraphs. The last paragraph of this chapter presents two examples of QALY modeling, one example at the individual level and one at the societal level.

I: The QALY Model Stage

First, a choice has to be made for a particular QALY model. This choice will not be based on theoretical considerations only, but on practical considerations as well. Each model arrives at a summary value for the combined health-status/survival outcome, but the models differ in

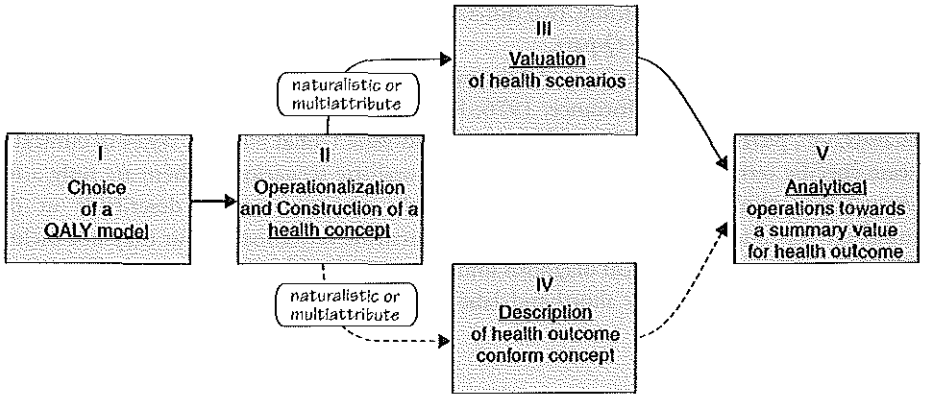


Figure 1
The five-stage approach to obtain summary values for health outcome

the degree to which the two basic elements, health status and survival, are valued either separately or jointly, and used in the computational procedures to arrive at QALYs.

Basically, four aggregation levels of a model can be distinguished. The lowest level of aggregation implies separate valuations of the domains of the health-status concept (e.g., physical, psychological and social functioning), and separate valuation of survival. This procedure requires complex computations. The highest level of aggregation implies the valuation of an entire (health) outcome tree (multiple lifetime health profiles), which is a complex task. However, complex computations are avoided in this case. All so-called QALY models have in common that they integrate information about valuations of health status (Q), life years (L) and eventually probability/uncertainty in a particular manner, and that a model-specific part of this process rests on individual valuation procedures, either by a panel or by the patient him/herself.

The L-part will not be further discussed in this introductory chapter, unless stated otherwise. We assume that the valuation of the years alive is equal to the numerical value of the (expected) life years. The formal notation for this assumption is: $U(L)=L$, where $U(...)$ stands for the valuation of a QALY element.

For the moment we will not consider the uncertainty element, although this is obviously an important aspect in individual decision making. In the context of a decision tree future health is reflected by branches with their probabilities of occurrence. The uncertainty element will be introduced at the end of this chapter when discussing the standard QALY model for the purpose of individual (patient) decision making. Also the difference of societal versus individual use of the QALY model will then be discussed. Implications and conclusions about these two different applications will be addressed in Chapter 10.

II: The Health Concept Stage

At this stage, a concept of health has to be defined, that generally comprises the element Q and eventually L. For the Q-part, it is generally agreed that a comprehensive concept should be chosen, covering at least the physical, psychological and social domains (Torrance, 1986; Froberg & Kane, 1989a). Basically, two different strategies (Froberg & Kane, 1989a) can be distinguished to construct the Q-part.

Holistic or naturalistic scenarios — All the early pioneering work on the valuation of health states (Patrick *et al.*, 1973b; Torrance, 1976; Rosser & Kind, 1978) was primarily based on scenarios, i.e., phrased descriptions of particular end-state ("final") health states in clinical or disease-specific terms. These are referred to as naturalistic scenarios. If there are only a few homogeneous clinical outcomes of interest, each outcome can be described[‡] in this naturalistic way and can be valued separately. Naturalistic scenarios may or may not include explicit statements of the clinical diagnosis underlying the health state described. They can be presented in several ways, by shorthand labels or brief narrations, even with multimedia support.

‡ We have referred to written scenarios, although other modes of presentation are conceivable, e.g., multimedia presentations.

Multiattribute scenarios — The main feature of multiattribute scenarios is that health status is described by a fixed, generic health classification rather than by a set of disease specific descriptions. Such a classification distinguishes several attributes. Each attribute represents a particular domain associated with health status. The attribute usually is subdivided into a number of levels, enabling a person's score on that attribute to be represented by a single item.

III: The Valuation Stage

Assigning numerical values to health-state scenarios is the third stage of the process of arriving at summary values for health outcomes. At the Valuation Stage, five choices are relevant. First, it is necessary to choose a specific valuation method as well as determining the precision of the scale measurements. The second choice involves deciding which subjects should perform the valuation task: e.g., physicians, experts, patients, panels or the general population. Third, the stimulus mode presentation has to be selected, e.g., multimedia presentation, orally, telephone, booklet. Fourth, the mode of response has to be selected: pencil-and-paper, face-to-face, telephone, multimedia or by the use of a collective voting system (see: Chapter 3 and Chapter 5). Finally, a decision has to be made regarding which scenarios or how many scenarios should be constructed and valued. For naturalistic scenarios, all health states of interest will normally be valued. for the valuation of (a sample of) multiattribute scenarios, certain requirements have to be fulfilled in case they are used to estimate a value function for the health classification. In the separate paragraph on the Valuation Stage (see: p. 34), the first choice is described in more detail. The methodology to develop a model based on the valuation of a set of multiattribute scenarios is summarized in the detailed section on the Analytical Stage (see: p. 38).

IV: The Description Stage

At this stage, the health of the target population is measured using an appropriate device for the *Q*-part and the *L*-part. Classifications in use for health status (*Q*) usually involve standard questionnaires for self-assessment, e.g., the EuroQol instrument (Brooks, 1996). Several methodological problems have been identified at the Description Stage. For instance, the questionnaire should explicitly specify the time period on which the self-assessment of the subject's health status should be based. In some studies, descriptions of health status are not obtained from the respondent but through assessments by experts, physicians or proxies (e.g., parents).

Knowledge about the psychometric properties of health-status classifications is limited, especially about their relationship to conventional quality-of-life measures (Staquet *et al.*, 1998). Another issue is the appropriate method of representation in the case of a naturalistic approach.

No further mention will be made about the description of health status in this thesis. For a detailed discussion on the descriptive measurement of health status (quality-of-life), see Essink-Bot (1995).

Table 1

Overview of the five methodological stages related to the process of obtaining summary values for health outcome

Stage I QALY Model	Stage II Health Concept	Stage III Valuation	Stage IV Health Description	Stage V Analytical Operations
Analytical Design	**Naturalistic Measurement**	**Tradeoff Method**	**Naturalistic Scenario***	**1. Value Function**
I. fully decomposed		standard-gamble		*(data of Stage III)*
II. partially decomposed	**Multiattribute Measurement**	time-tradeoff	**Multiattribute Scenario**	functional measurement**
	Rosser & Kind Index	willingness-to-pay		multiattribute utility model***
Direct Assessment	Quality of Well-Being scale			multiple classification analysis
III. single profile	Health Utilities Index I–III	**Magnitude Estimation Method**		multiple regression analysis
IV. multiple profiles	EuroQol-5D	magnitude estimation		
		equivalence technique		paired comparisons****
		person-tradeoff		conjoint measurement****
				unfolding****
		Rating Method		
		rating scale		**2. Descriptive Analysis**
		visual analogue scale		*(data of Stage IV)*
		Preference Method		**3. Estimation Values Actual Health States**
		binary choice		*(Q-data of Stage III; IV)*
		pick one of N		
		rank order		**4. Summary Values (QALYs)**

* Condensed or narrative format; these type of scenarios are possible with or without disease labels.
** Based on analysis of variance.
*** Model without error component.
**** Model based on preference method data.

V: The Analytical Stage

The computational procedure required to arrive at the summary measure (in QALY units) takes place at the Analytical Stage. Especially for the valuation of multiattribute health scenarios, a considerable amount of analytical procedures is demanded. In that case, usually only a small subsample of health states have been valued at the Valuation Stage, whereas the estimation of a reliable and valid value function to predict the non-valued health states is cumbersome. Prior to the estimation of a value function, experimental studies focusing on the determination of factors other than the stimuli of the multiattribute scenarios are indispensable. The associated analysis is also included at the Analytical Stage. Subsequently, the descriptive classifications obtained at the Description Stage are quantified by applying the value function estimated at the Analytical Stage. The separate elements are then combined into one summary value.

In the following four paragraphs we will discuss in more detail the QALY Model Stage (I), the Health Concept Stage (II), the Valuation Stage (III) and the Analytical Stage (V).

Table 1 shows an overview of the five stages and their components. The many terms and names in the table will be discussed in the following sections.

Stage I: The Choice of a QALY Model

The concept of Quality-Adjusted Life Years (QALYs) as a summary value of health status and survival in the context of health outcome assessment was developed in the early 1970s. The concept first appeared, as "function-years", in a compelling paper by Fanshel & Bush (1970). Obviously unaware of this publication, Grogono & Woodgate (1971) published a similar approach to evoke indices to measure health. They specified health according to ten domains and arrived at "health-years" after a simple weighting algorithm. Six years later in 1977, Weinstein & Stason published their landmark article in which they described the QALY (gained) as the appropriate measure of effectiveness of treatments. The definition of the QALY, as stated in their article, was:

> A health-status index is essentially a weighting scheme: each definable health status, ranging from death (...) to full health (...) is assigned a weight zero to one, and the number of years spent at a given health status, Y_s, is multiplied by the corresponding weight, λ_s, to yield a number, $\lambda_s Y_s$, that might be thought of as an equivalent number of years with full health – a number of quality-adjusted life years (QALYs).

The authors asserted that the source of these weights is ultimately subjective, but this fact does not exclude "societal consensus". The QALY model in their view encompasses two elements. Mortality data is introduced in the global QALY model as the Y_s (survival, e.g., life years) parameter, whereas morbidity data is introduced by the λ_s parameter. Parameter λ_s is the parameter that has to be evoked by the valuation of the health state under study. Note that

Weinstein and Stason assumed that the numerical quantity of L is equal to its value, which is a subjective assumption. Moreover, simple multiplication of λ_s with Y_s, and weighting with the probability of occurrence of $\lambda_s Y_s$ is assumed to reflect the overall value (utility, preference). (The parameters Y and λ are in this thesis indicated respectively as L and Q respectively).

The standard decision model utilized by scientists in analyzing decision problems is derived from Expected Utility Theory (EUT), and the standard graphical tool is the decision tree (considered to be equally applicable at all decision levels). The theory and its methods of measurement were developed in the field of economics and operation research as a normative model for individual decision making under uncertainty. According to this theory, utilities are (cardinal) numbers representing the strength of preferences for particular outcomes. The theory was developed by the mathematician von Neumann and the economist Morgenstern. They developed an axiomatic theory of rational decision making under uncertainty in 1944. This theory is normative or prescriptive, that is, it prescribes how a rational individual should make decisions when faced with *uncertain outcomes* to increase his/her welfare in the most efficient way (von Neumann & Morgenstern, 1953). EUT has been widely applied in operations research for business, government, policy, health care, and many other fields for more than three decades. Medical decision science has adopted EUT and its accompanying methodology to derive preference measures (utilities) of patients for medical treatment decisions (individual level).

Pliskin *et al.* (1980) were the first to present a formal model that comprised two distinct elements: health status (Q) and life years (L), that arrived at a consistent quantitative measure. Their QALY model combined the elements of Weinstein and Stason with the formal aspects of EUT. The model was developed for outcomes described by a specific number of years in a constant (chronic) health state. According to this model, the value (formally called utility) of the outcome can be described as follows: pairs (Q, L) where L is the survival duration and Q is *permanent* health status experienced *during* the years of survival:

$$U(Q,L) = bL^rU(Q), \qquad (1)$$

where b is a scaling constant and r is a risk posture (c.q., attitude) parameter. $U(Q)$ is a measure of health status adjusted for its quality between 0 and 1. It should be noted that the value of L in this model is assumed to be equal to the numerical quantity of L^r, formally stated as $U(L)=L^r$. The axioms of EUT and the empirical work of Keeny & Raiffa (1976) point out that there are three conditions to be fulfilled in order to compute, in this case, QALYs that will represent a valid utility function for health with a constant quality of health over time.

The first condition is that the valuation of the two elements, Q and L, must be *mutually independent*. This means that the proportion of the utility of full health (Q^*) for any health state is the same for all time horizons (Figure 2a). The second condition is *constant proportional tradeoff*.[§] This implies that the proportion of remaining life that one would be

§ Time preference (i.e., discounting) is not explicitly incorporated in this model. In order to incorporate discounting into the Pliskin model, the constant proportional tradeoff has to be assumed with respect to

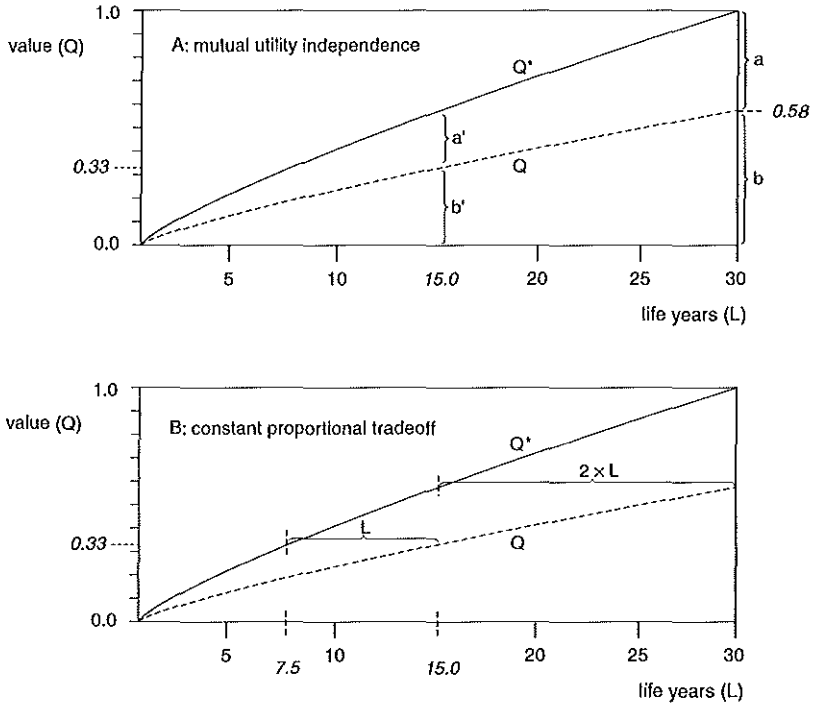


Figure 2
The assumption of constant proportional risk posture ($r = 0.8$) imposing,
A: mutual utility independence (ratio $a/b = a'/b'$) and,
B: constant proportional tradeoff (the time interval, here arbitrarily chosen at 30 years, at which health states Q^* and Q have equal utilities is proportional; $L/7.5 = 2 \times L/15$)

willing to trade off for a specified quality improvement is independent of the amount of remaining life (Figure 2b).‖ This later condition implies that the reverse also holds, e.g., the evaluation of life years is the same for all health states.

If both the independence assumption and the proportional tradeoff assumptions hold, the model as defined by Pliskin *et al.* is valid and the shape of such a utility function is known as constant proportional risk posture over life years. Under these conditions, the imputation of the proposed r is still required. On some occasions r may be known, but in most circumstances this parameter has to be estimated. The interpretation and estimation of r is not easy as its inclusion seems to rest on normative rather than data-descriptive considerations (Appendix A will show it can be safely omitted). For this reason and because the model is only

discounted life years rather than with respect to life years.

‖ The second condition is the more general, as it reduces to the first if $U(Q)$ of the period to value and $U(Q)$ of remaining life are the same.

applicable for very restrictive health situations (e.g., permanent health status during the years of survival), this QALY model has seldom been used in actual applications.

Studies that presume that the shape of the utility function over life years is linear ($r = 1$) avoid these difficulties but in fact assume *risk neutrality regarding life years*. If this condition holds, the two lines in Figure 2a en Figure 2b are linear.

We will now further elaborate on the different available approaches to evaluate health status (and survival) and associated computations to obtain QALY measures. Four QALY models are distinguished according to their level of aggregation. These four models incorporate an increasing number of elements in the empirical valuation task, and a decreasing number of analytical operations.

Schematic representations for these models are depicted in Figure 3. Solid lines indicate a decomposed element of health. It can be seen that model I is fully decomposed; the attributes (A) of the Q element and the element L are separate. In contrast, Q and L are not represented as separate elements in model III. To facilitate the discussion of the models, definitions are given for two specific terms.

- A *QALY-ple* is a discrete period, uniquely defined by:
  1) constant health status (Q), optionally subdivided into a set of attributes (A) cf. some classification;
  2) duration (L);
  3) a censoring end-point (C; not discussed in this introduction, see: Appendix A).
  The QALY-ple is assumed to be the smallest homogeneous unit of health outcome.
- A *health profile* is defined as a combination of QALY-ples; an *episodic health profile* is defined as a representation of a health profile comprising two or more QALY-ples; a *lifetime health profile* is a health profile, usually consisting of multiple QALY-ples, the last QALY-ple ending with death. The Pliskin model therefore is a lifetime health profile consisting of one QALY-ple (in Figure 3: QALY model II, lifetime health profile A). Combinations of QALY-ples are represented in bold capital typeset, e.g., $U(\mathbf{Q}, \mathbf{L})$.

An extended mathematical description of the four QALY models and an elaboration of the underlying assumptions is presented in Appendix A (p. 183).

Model I: Fully Decomposed

Health states in the Fully Decomposed QALY Model are not valued as a whole but are decomposed into the distinct attributes that comprise a health-state concept. Each attribute is separately valued. For both the health-status element and the survival element this model simply assumes that they may be summed.

Figure 4 shows two modes. The second mode is a specific form of the first mode based on one additional assumption. If health status is assumed to be constant for a whole lifetime (e.g., chronic diseases), the Fully Decomposed QALY Model requires the assessment of scenarios with fixed time periods (e.g., one year) for each attribute of the health concept (Figure 4, Example A). For non-constant health profiles, not only does each attribute have to be assessed, but the valuation task has to be performed for each separate health status in the health profile (Figure 4, Example B).

Model II: Partially Decomposed

The model of Pliskin *et al.* (1) was derived for health profiles with constant health status over time (Figure 5, Example A). It is unclear how to apply this model if the quality varies over time, which is more usual in cases of impaired health status. Diseases with variable health status form a major part of the disease spectrum and it is essential to find ways of dealing with such conditions.

The common model that deals with varying health states is based on the rather simple assumption that a time period of different health states may be treated as a summation (additive assumption) of smaller periods, each with constant Q.

It should be noticed that the QALY model of Pliskin *et al.* is different from the Partially Decomposed QALY Model as presented here. Pliskin's model assumes certain specific conditions (see: p. 22) and adds a risk parameter (r) to L to determine $U(L)$.

As can be seen in Figure 5, the Partially Decomposed QALY Model assesses the set of health-status attributes (A) integrally. Instead of separate valuations for the distinct empirical attributes implied by the Fully Decomposed QALY Model, here all the attributes (q_1, q_2, ..., q_A) are valued as one. For the permanent constant health profile situation, Partially

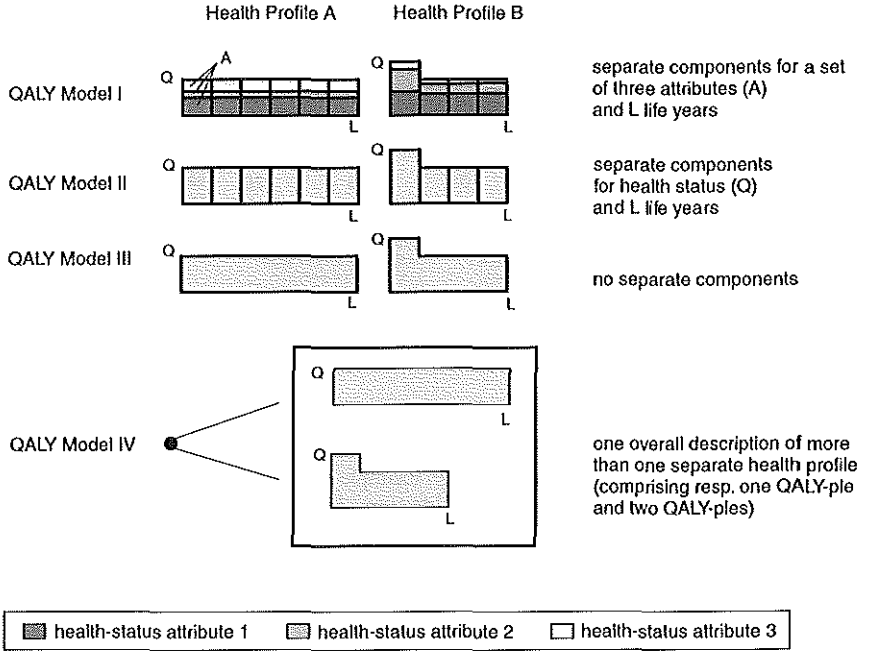


Figure 3
Conceptual representation of the four different QALY models (I – IV) for a situation with two different lifetime health profiles, assuming $U(L) = L$; the solid lines indicate that the elements are explicitly and separately valued

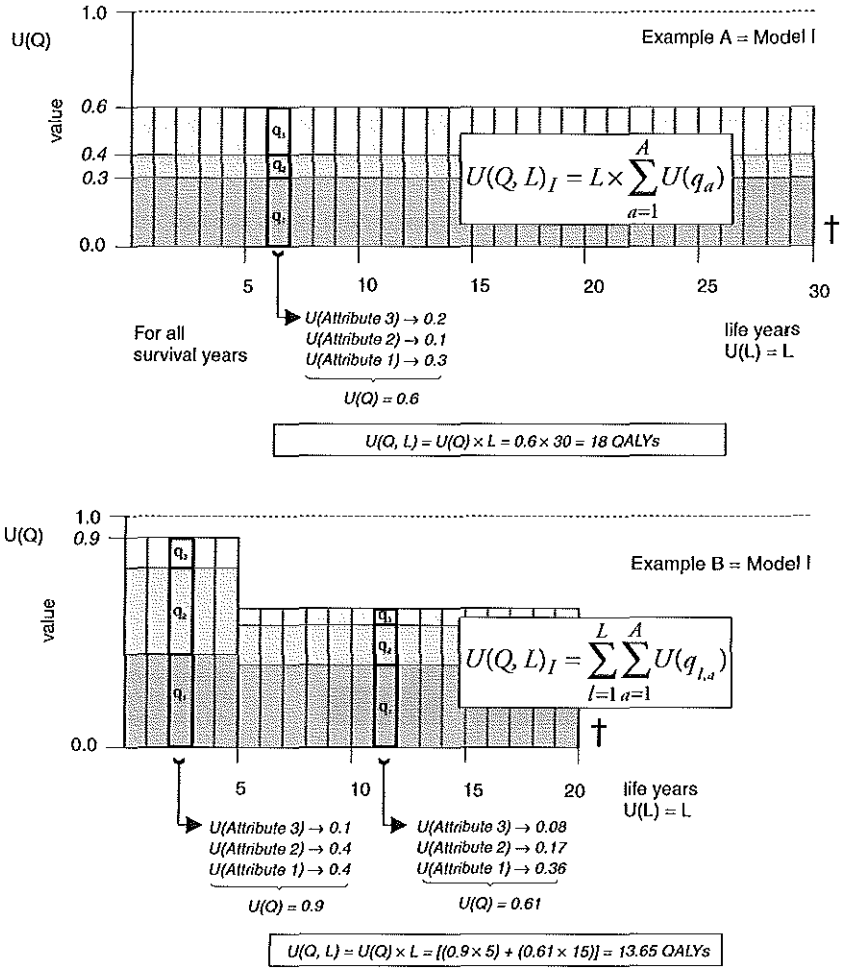


Figure 4

MODEL I — Conceptualization of the Fully Decomposed QALY Model, valuations for each separate health-state attribute (here 3 attributes), *U(Q)*; Example A: for a fixed time period (here one year) multiplied by their corresponding number of life years (*L*) for a chronic condition; Example B: the same as Example A, but for two different health states in two consecutive periods

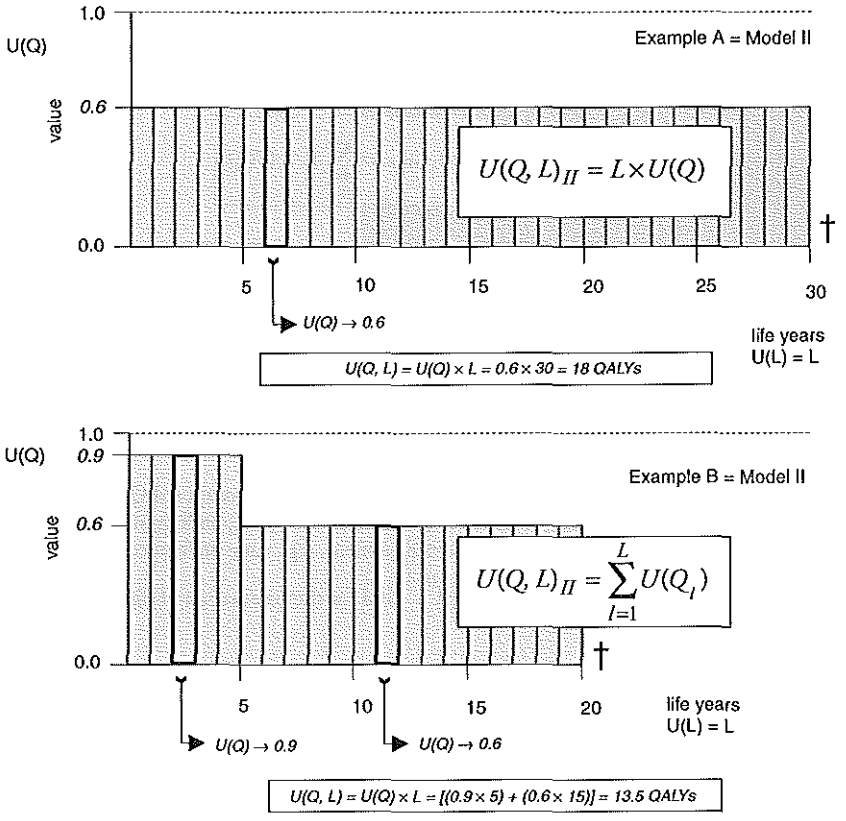


Figure 5
MODEL II — Conceptualization of the Partially Decomposed QALY Model; valuations, *U(Q)*; Example A: for a fixed time period (here one year) multiplied by their corresponding number of life years (*L*), Example B: the same as Example A, but for two different health states in two consecutive periods

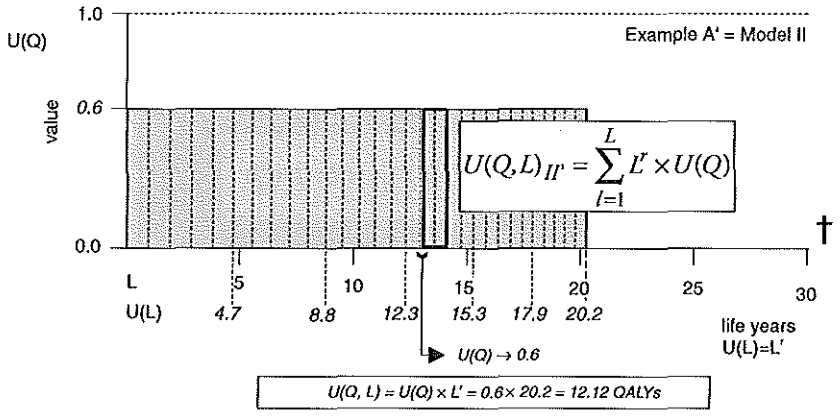


Figure 6

MODEL II with time preference — Conceptualization of the Partially Decomposed QALY Model; Example A': for a fixed time period (here one year) multiplied by the corresponding number of life years (L); depicted is the implicit appraisal of the survival element, e.g., time preference, which is in this example for each specific life year: $U(L)_n = \left(\frac{1}{1+r}\right)^{n-1}$ and for the summation of discounted life years, $\sum_{n=1}^{N} U(L)_n = \frac{1-\left(\frac{1}{1+r}\right)^n}{1-\left(\frac{1}{1+r}\right)}$ (n = year: here 30 years; r = discounting factor, assumed to be constant over years: here 0.03)

Decomposed QALY Models require first to value Q (naturalistic or multiattribute) for a defined L (for valuation purposes usually one year), followed by multiplication with L (Figure 5, Example A). Health profiles of non-constant situations require valuations of health status for each separate QALY-ple of the health profile (Figure 5, Example B). Another possibility is the valuation of Q, multiplied not with L however, but with L adjusted for a specific assumed time preference (Figure 6).

Model III: Single Profile

Model III is based on the value of entire health profiles (Figure 7). Hence, it assumes a higher empirical aggregation level compared to the previous two QALY models. If health status varies, it is not required to value each distinguishable health state separately (Figure 7, Example B).

No further computations are required for this QALY model if only one specific health outcome is observed. All effort has to be dedicated to the careful presentation of the information of the two elements, Q and L. The valuation task consists of the assessment of, in the most simple case, one naturalistic scenario and one that can be mathematically described as $U(Q, L)$ for the chronic state or $U(Q, L)$ for states of multiple QALY-ples. One of the major

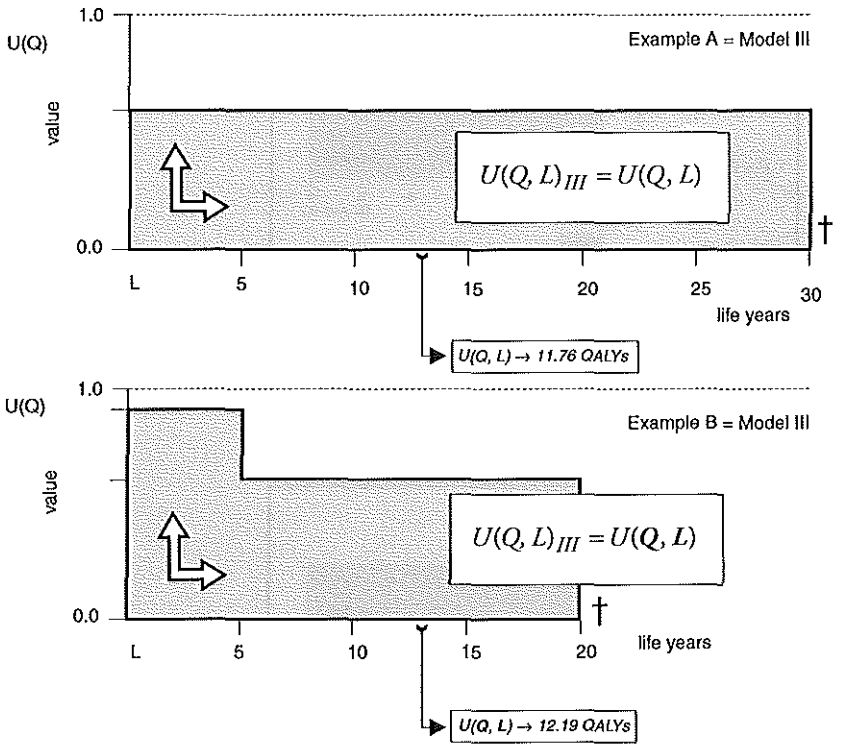


Figure 7
MODEL III — Conceptualization of the Single Profile QALY Model; integral valuation of the total health profile (no further computations); Example A: for a constant health profile (here 30 years), *U(Q, L)*, Example B: The same as Example A, but for a health profile with varying health status during a period of 20 years, *U(Q, L)*, (no further computations, indicated by the arrow in the box)

advantages of such a direct (also called holistic) assessment approach is the incorporated assessment of *L*, instead of assuming *U(L)* = *L*.

The idea of valuing health profiles rather than health states is a logical step, first touched upon by Mehrez and Gafni (1989) in their Healthy Years Equivalence (HYE) procedure, which is one of the possible empirical valuation procedures to deal with so-called Single Profile QALY Models.

Model IV: Multiple Profiles

The highest aggregation level is achieved if a full array of outcomes is simultaneously valued. This QALY model may be relevant for any case with two or more different possible health outcomes (Figure 8). As elaborated in Appendix A, we may distinguish a societal variant and an individual variant. For individual clinical decision making (by patients), this valuation

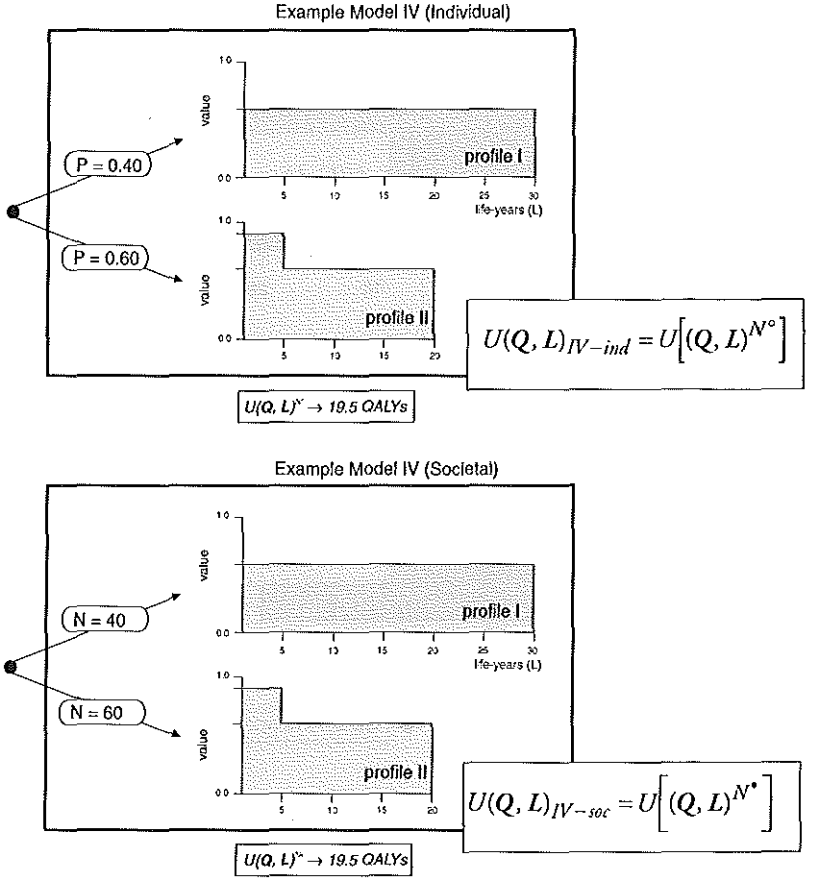


Figure 8
MODEL IV — Conceptualization of the Multiple Profile QALY Model; integral valuation of multiple health profiles (no further computations) with the inclusion of probability information (individual level) or distribution information (societal level)

approach encompasses the integral assessment of the three relevant elements, Q, L and the probability (in Figure 8: P) associated with each health profile. This model does not necessarily assume that $U(L) = L$, nor that the simple probability weighting function is valid, although empirical studies may yield results that support these two common assumptions. All three element are assessed integrally, where Q and L are genuine elements, and probability modifies their values.

In Appendix A we will show that for societal based QALY models we can incorporate distribution (justice or equity) considerations in a way similar to the incorporation of probability considerations in the individual case. The formal notation in this appendix is

worked out for the individual and the societal based QALY model, respectively as: $U(Q, L)_{IV\text{-}ind} = U[(Q, L)^{N^o}]$ and, $U(Q, L)_{IV\text{-}soc} = U[(Q, L)^{N\bullet}]$.

Stage II: The Operationalization and Construction of a Health Concept

After a specific concept of health has been defined, the next step is its operationalization. Because we assume $U(L) = L$ we focus on the specification of the health status. Health-status values need to be based on a broad definition of health that incorporates at least the physical, psychological and social domains. We will focus on the two existing approaches to construct health scenarios for health status only.

Naturalistic scenarios

All the pioneering work on the valuation of health states has been based on narrative descriptions of particular health states, in terms of symptoms or clinically relevant disease states with or without the explicit mentioning of the particular clinical diagnosis. If there are only a few health-state outcomes of interest, each health state can be described separately in this way and the value for each can be obtained directly by the use of any valuation method. For those situations where a considerable number of health states are relevant, this direct strategy becomes cumbersome.

For clinical decision making, which is usually an individual process, scenarios have to be understood as if they are "own" (possible future) health state of the subject him/herself. Naturalistic scenarios are suited for such purposes. But how should we describe, in a comprehensive and yet unbiased manner, a particular dysfunctional health state (for example "angina pectoris" or "cardiovascular incident") to a (healthy) individual who has no experience

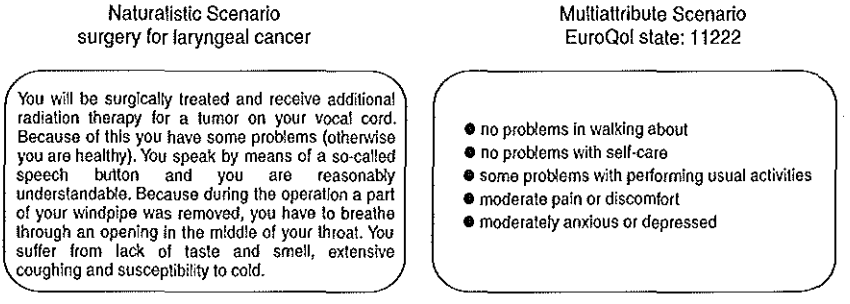


Figure 9
An example of a naturalistic narrative health state scenario versus its multiattribute based shorthand counterpart for the same health state

of such a condition? In fact, even patients usually have very limited personal experience with most of the (possible) health conditions associated with their disease.

Naturalistic scenarios including clinical diagnosis, for example, were used in a recent population-oriented study carried out by the Harvard University & the World Health Organization project commissioned by the World Bank (Murray & Lopez, 1996). A large set of diseases and impairments were valued by various expert panels. Several options are available to construct naturalistic scenarios. The conventional narrative or condensed forms are possible, but such scenarios may also solely consist of disease labels. Even combination of narrative or condensed forms with disease labels are permitted.

Depending on the problem involved, naturalistic scenarios can be presented as life-long scenarios, including variability of the health status or they can be presented as lasting for a defined period (e.g., one year).

Multiattribute scenarios

Several classification systems have been developed to construct health-state scenarios founded on a specific decomposed health concept. For example, health state scenarios based on the EuroQol system are constructed from five attributes. Each attribute can be varied on three levels (Brooks, 1996). The five attributes of this health system are: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Their levels are expressed as: no problems, some moderate problems and extreme problems. Assuming independency 3^5 (243) different health states are defined by the EuroQol classification. Every situation (health state) is assumed to be approximately covered by one of the 243 generic scenarios. The advantage of multiattribute scenarios is that if a suitable subset of health states is empirically valued, the prediction of the remaining states ("the valuation space") can be estimated by interpolating some multiattribute function. Simple multiattribute classification systems permit complete empirical valuation of all possible health states (Rosser & Kind, 1978; van Busschbach *et al.*, 1997), but usually this is not feasible.

Other classification systems are: the Health State Classification System, later renamed as the Health Utility Index Mark I (Torrance *et al.*, 1982), and its successors (HUI-Mark II and III), developed at the McMaster University (Feeny *et al.*, 1994; Torrance *et al.*, 1995; Torrance *et al.*, 1996). One of the first classification systems was the Quality of Well-Being scale (Kaplan *et al.*, 1976; Kaplan & Anderson, 1988). An example of a naturalistic scenario as opposed to its counterpart, a multiattribute scenario, is shown in Figure 9, whereas the relation between the number of attributes and number of levels of the classification systems is depited in Figure 10.

It should be noted that life years, the L-part of the health concept, can also be operationalized in different ways. As already mentioned, the usual approach is the metric measurement or estimation of survival [$U(L) = L$]. An approach where L is assessed by qualitative expressions of survival analogous to Q is conceivable.[5]

5 For example: "soon death", "average life span", "terminal stage".

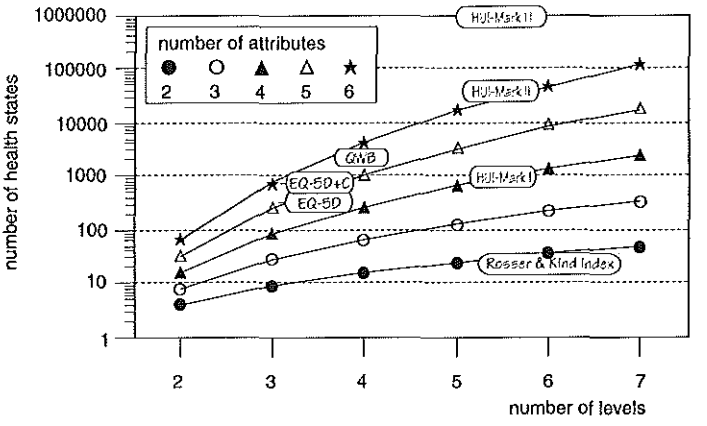


Figure 10

The number of scenarios of health states that can be constructed based on combinations of numbers of attributes (n) and number of levels (p) (some frequently applied multiattribute classification systems are depicted at their appropriate position: Rosser & Kind index, 32 health states; EuroQol-5D, 243; EuroQol-5D+C, 729; QWB, 1215; HSCS/HUI-Mark I, 960; HUI-Mark II, 24.000; HUI-Mark III, 972.000); the number of *dominant* pairs of health states for any classification system, with an equal number of levels for each attribute, is given by the formula $\left[\frac{p(p+1)}{2}\right]^n - p^n$

Any classification system is a compromise between, on one hand, the comprehensiveness (attributes) and refinement (levels) of the classification, and, on the other hand, the feasibility and the precision of the evoked assessments. Methodologically, this tradeoff is about the validity and the reliability of classification systems. A recurrent question is whether the number of attributes of an existing classification system should be extended or reduced. Similarly, the number of levels within an attribute may be questioned. In particular, an increase in the number of levels, increases the sensitivity at the Descriptive Stage. We provide an example of such a study on the validity of a multiattribute classification system in this thesis (Chapter 6). For the standard EuroQol classification system (EQ-5D) comprising five attributes each with three levels, we constructed an extended version (EQ-5D+C) by adding a 6th attribute (cognitive functioning). By collecting valuations for both versions of this instrument, we were able to examine the consequences of extending this classification system (see: Chapter 6).

Methodological considerations are also relevant during the selection of the number of attributes in relation to their number of levels. The benefit of a comprehensive health-status classification system with small intervals between the levels is the ability to differentiate more accurately between health states. However, the increasing number of health states that have to valued and the complexity of the associated analysis (Analytical Stage) is a disadvantage when trying to arrive at a valid classification system.

Stage III: The Valuation of Health Scenarios

At this stage several choices are relevant. The choice of an appropriate valuation method is probably the most controversial, even though a considerable amount of evidence may be available to support a particular choice. Several methods have been proposed to elicit values for health states, usually involving a state of fixed duration and fixed health status (decomposed model oriented). The most frequently used valuation methods will be explained. Three of these methods are so-called tradeoff methods. Here the valuation task consists of sacrificing a specific element (probability, survival, money) in favor of a defined improvement in health. The other three methods are more orientated towards the direct numerical valuation of health states.

Valuation methods

Tradeoff methods

Standard-Gamble (SG) — The SG concept is derived from the von Neumann-Morgenstern (vNM) utility gamble. The method as it is currently operationalized is an approach that aims to achieve indifference between two outcomes under conditions of uncertainty. The respondent is presented with two alternatives, and asked to select the preferred one.[¥] One alternative offers the respondent a specified certain outcome, while the other alternative offers a gamble with mutually exclusive probabilities for two reference outcomes. Conventionally, SG is operationalized as a choice between A, the certainty of being stationary in a specific lifelong impaired health state (the state to be valued), or B, the uncertain result of an intervention, for example a surgical procedure, with two reference outcomes. These reference outcomes are: a probability (p) of instantaneous and lasting improvement to perfect health, or a probability ($1-p$) that the operation will fail, resulting in immediate death. By varying the p-level of the uncertainty outcome (iterative procedure), the point of indifference between the two alternatives A and B is determined. We will call the numerical expression of a valuation method (also called: elicitation method) the "calibrator". For SG this is p. By combining the probability values with the utility values of the reference outcomes, the utility of the stationary state is established. The method fails if the state to be valued falls out of the range covered by the two reference outcomes (for example, if in the conventional operationalization a state is considered by the respondent to be worse than death).

SG is the only method where, due to its specific operationalization, health states are valued under risk (p, $1-p$). The numerical quantities evoked by SG are therefore called vNM-*utilities*. All the numerical quantities evoked by other elicitation methods are referred to as *values*. It is

¥ Hershey *et al.* (1982) distinguished four different methods for constructing vNM utility functions, namely:
1. the probability equivalence, where an indifferent level for p is elicited;
2. the certainty equivalence method, where the health state scenario is altered until indifference is reached;
3. the gain equivalence method, where the probabilistic outcome of the best health state is elicited, and
4. the loss equivalence method, where the probabilistic outcome of the worse health state is elicited.
The standard-gamble method in this terminology is equal to the probability equivalence method.

for this reason that some scientists assert the opinion that only the figures elicited using the SG method yield true utility functions. SG is often regarded as the right method to elicit values for clinical decision analysis, as SG is thought to be based on EUT and the valuation task includes aspects of uncertainty (e.g., risk, p and $1-p$)

Recently, Richardson (1994) has pointed to a flaw in this method, in particular when applied in clinical decision making (individual level). He has stated that risk is introduced by the method as part of the technique (p), but this p-value is generally not the same as the probability (risk) associated with the outcome of the particular medical intervention under study. Hence, method-uncertainty $\neq$ stimulus-uncertainty, and superiority claims based on this argument are invalid.

Time-Tradeoff (TTO) — The TTO method was developed by Torrance (1976) as a less complicated, conceptually different, alternative to SG. It is based on tradeoffs similar to those of SG, but the concept of probability is replaced by time, as the calibrator trades off survival and health status. The first alternative specifies a (suboptimal) health state with a given duration of, say, ten years. The competing alternative offers a better health status (conventionally optimal health) of shorter duration. The point of indifference is reached by varying the duration spent in perfect health. By combining the duration value and the values of the reference outcomes, the utility of the stationary state is established.

The flaw associated with the TTO method, as with the SG method, specifically relates to the operationalization of the method itself as mentioned above. In particular, the so-called time preference (method-related and stimulus related) interacts with the values elicited by this method. This issue will be discussed more extensively in another context (see: Chapter 6). For references about methodological consequences related to TTO see for instance Gafni and Torrance (1984; Gafni, 1995), Johannesson *et al.* (1994) and Dolan & Gudex (1995).

Willingness-To-Pay (WTP) — From an economic point of view, WTP can be considered as the superior monetary quantification of non-monetary aspects of disease (Thompson *et al.*, 1982; Thompson *et al.*, 1984; Gafni, 1991; Gold *et al.*, 1996; O'Brien & Gafni, 1996). Essentially a subject is asked how much he/she would be willing to pay (either an insurance premium or a direct amount of money) to obtain a well-defined gain in health. One of the earliest applications of the WTP method for assessment of health consisted of the insurance-like question: "What percent of your income would you be willing to pay on a regular basis for a complete cure of your arthritis?" (Thompson, 1986). However, for several reasons the method has not very often been used. A principle objection has been the potential of unwanted effects as the value of money depends on one's socio-economic class, although it is unclear how so-called distribution effects (Gold *et al.*, 1996) will occur *across* diseases. An equally important drawback is uncertainty about the correct operationalization. For example, should the method take the form of an insurance question or a "pay-direct" question, either from the patients' perspective or from the perspective of the healthy person (see also: Chapter 4)?

Magnitude estimation methods

Magnitude Estimation (ME) — This method has initially been used widely in the field of psychophysics. The core of the technique is that a subject is asked to provide a subjective ratio by assigning numbers to a certain stimulus (for example; intensity of light or sound) in

proportion to the number that has been assigned to a reference stimulus (Stevens, 1957). In the context of the valuation of health states, scenarios have been used as stimulus. Whereas the tradeoff methods SG, TTO and WTP are characterized by tradeoffs, respectively probabilities, life years and money versus improvement of health, magnitude estimation is focused on making a comparable subjective quantification for *identical entities.*

The *Equivalence Technique*, first introduced by Patrick *et al.* (1973a), is an adaptation of magnitude estimation. It has been applied in various forms, but the common underlying task is, for example, to decide how many people in health state A are equivalent to a specified number of people in health state B. One of the earliest applications of the equivalence technique was used by Rosser & Kind (1978) to construct their health-status matrix. Recently this method and extended versions have been advocated by Nord (1995), although under another name, i.e., person-tradeoff (PTO). The method has also been used in a large study by Murray & Lopez (1996). In this study a subject is offered two alternatives. Alternative A is to extend life for x individuals in normal health and alternative B is to extend life for y individuals in health state i. Y is varied until the respondent is indifferent to the choice between the two alternatives at which point the preference for state i is x/y. Modified forms of PTO have been devised where subjects are asked to tradeoff restoring health to x individuals in health state i versus restoring health to y individuals in health state j (Murray & Lopez, 1994; Nord, 1995). This extended version of PTO is based on a combination of the equivalence technique (number of subjects) and a tradeoff task (restoring health).

PTO and its extended variant has been recently advocated for its appropriateness for eliciting societal values for health status. This is because the main feature of the method involves a tradeoff between subjects instead of a tradeoff between certainty versus uncertainty (SG) or life years (TTO). Richardson's argument (1994) in regard to SG however, also applies to PTO, i.e., the (dis)utility of numbers (subjects) in the PTO task on Q, has little, if anything, to do with the societal disequilibrium associated with mutually exclusive allocation choices. Furthermore, it should be noted that whereas ME is based on expressing health outcomes as subjective ratios, PTO could be viewed as a more confusing method, partially based on eliciting ratios for the two health states to be assessed and partially a tradeoff method using the tradeoff of subjects as units of measurement. Apart from the possible appropriateness of the PTO method in eliciting societal values, the cognitive task required from participants is complicated, probably forcing them to use simpler heuristics than the task assumes.

Rating methods

Rating Scale (RS) — For this popular method, respondents are simply asked to express their view by putting a mark on a line divided by predetermined intervals (such as a thermometer) with two descriptive anchors at both end-points. The anchor states are usually marked (but not by definition) as "healthy" and "dead", therefore encompassing the full range of health. The respondents' task is to locate one or several health-state scenarios on a scale between the two anchors in order of preference, in such a way that the intervals between the positions of the health states correspond to the differences in preference as perceived by the respondent. Formally we should make a distinction between RSs and visual analogue scales (VAS). The latter scale omits intervals and solely consist of a line with descriptive anchors at both ends.

This type of scale also has a long history and were initially referred to as for example "graphic rating" (Hayes & Patterson, 1921). One of the first applications of the VAS for measurement in medicine were presented by Aitken (1969) and by Zealley (1969).

The advantage of both the RS and the VAS is that they are easy to use. They also, however, have a number of drawbacks. For instance, it is questionable whether these scales yield values that possess interval characteristics. Moreover, compared to the tradeoff methods and magnitude estimation, a rating scale involves no tradeoff or choice, which leads to different response behavior (see: Chapter 4).

Preference methods

Preference Methods (PM) — Several PMs have been developed, and the results have generally been used for subsequent computations by the paired comparisons (PC) method. This PC method was also developed within the field of psychophysics and the method goes back at least as far as 1860, when psychophysicians carried out extensive experiments on the relationships between physical magnitudes (e.g., loudness, taste, duration [hence; small intervals, no life years], heaviness, visual distance) and their subjective (psychological) magnitudes (Stevens, 1957; Vroon, 1972; David, 1988). An essential condition for PC is that stimuli are largely subjective and without clear dominance from one over the other. In the conventional binary choice method (i.e., *paired* comparisons) health states (stimuli) are presented in pairs to judges, who must choose the preferred health state for all pairs. Analytical procedures pertaining to the PC choosing task are subsequently required to obtain *scaled* values for health states. According to our knowledge, preference methods such as *pick one of N* and *rank order* tasks together with appropriate psychometric scaling techniques, so far have not been used in the valuation of health outcomes.

Stage V: Analytical Operations Towards a Summary Value for Health Outcome

Introduction

After the valuation of health scenarios at the Valuation Stage, and after the measurement of the actual patient's health (or the health of individuals from a survey sample) at the Descriptive Stage, four related data operations are needed to obtain summary values for health outcomes, either at the individual or societal level. These four (analytical) operations (see: Figure 11) are:

V.1 The analysis of elicited values, based on multiattribute scenarios, to derive a *value function* from these values for a particular multiattribute health-status classification system (a);
or the *direct elicitation* of the individual preferences, based on naturalistic or multiattribute scenarios (b).

V.2 The analysis of the descriptive data (usually of patients) to derive a set of health-state scenarios in a way which adequately represents the patterns of health outcomes of the problem under study.

V.3 The estimation of the value for the actual health-state scenarios by inserting the descriptive data of the second operation into the value function of the operation V.1a.

V.4 The estimation of summary values (QALYs), using aggregation rules for the values under operation V.3.

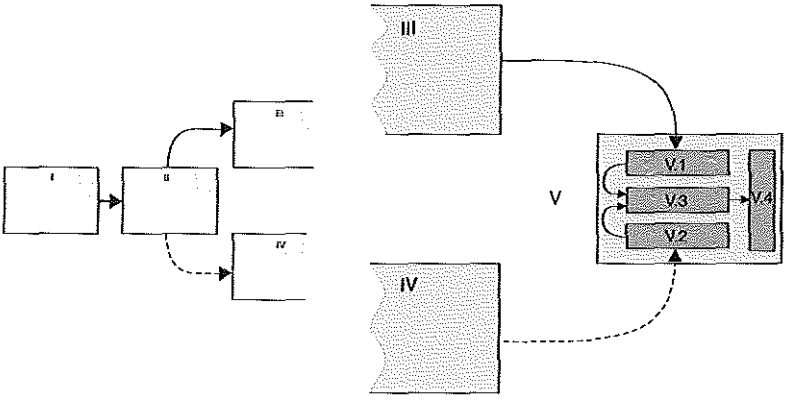


Figure 11
The four analytical operations at the Analytical Stage (V) of the approach to obtain summary values for health outcomes

Before we elaborate on these operations, two remarks should be made. First, the four operations differ according to whether a multiattribute scenario or a naturalistic scenario is chosen. Second, from the beginning of this chapter and from Appendix A, it follows that operations 3 and 4 are complementary. The QALY model chosen defines whether the health scenarios for operation 1 and 2 describe *Q* only, with estimation of *U(Q)* in operation 3 and the calculation of *U(Q, L)* in the operation 4 – which is the usual case in QALY modeling – or whether health scenarios in operations 1, 2 and 3 represent a profile with less complex computations in the last operation.

The following sections V.1 – V.4 assume a decomposed QALY model context, focusing on health status *(Q)*. Hence, the attributes that are analyzed in the first three operations are about this health element only. The survival element *(L)* can be added to this as a separate attribute, at least in principle. The four operation procedures will be discussed in due order.

V.1 — Analysis of the elicited values

The ultimate goal of the Valuation Stage is to estimate the contribution of the elements of the health scenarios (stimulus) to the value (response) of these scenarios. Ideally, health-state values are perfectly explained (and predicted) by the scenario stimulus, without error components. In practice, various known, yet "undesired" factors also contribute to the response.

The analysis of valuation data can be carried out in two ways. The first approach starts with a formal analysis of at least the known potential sources of variance (error) in order to decide whether the assumed analytical model for explanation is a valid choice and whether the reliability of the elicited values is sufficient. Step 2 of this first approach represents the usual way that in principle consists of specifying a value function. The second approach refrains from such a formal analysis of the properties of the data and from the testing of the existence of known sources of variance, apart from the assumed effect of the health scenario-stimulus.

The major non-health related factors in step 1 of this first approach are the valuation methods (see: Chapter 4), the respondent's characteristics (see: Chapters 4, 7 and 9), the stimulus mode and the response mode (see: Chapter 3). To disentangle these effects, experimental studies are necessary which combine the valuation of (a predefined set of) health scenarios with non-health stimulus factors based on some kind of a factorial design. As will be shown in Chapter 3 and 4, analysis of variance techniques (G-theory: Streiner & Norman, 1995) are particularly suitable to estimate quantitatively the contribution, if any, of non-health related factors.

To perform an effective analysis the (say H) health states of all the (say R) respondents' responses are used, taking into account the fact that the H observations of a particular individual are not independent. After analysis of the H × R data, a normative judgment is required to decide whether the contributions of non-health factors exceed "normal" variation and whether separate attention should be paid to such a factor in the second step. The typical analysis strategy followed in step 2 in the area of health-state modeling is to aggregate the data of the H health-states across R respondents to mean or median values. Hence, the H × R data is reduced to a single data vector with H "measures". The aggregated data sets are subsequently analyzed by one form of regression analysis to estimate the contributions of the various

components of the health description to the (average) health-state values (Torrance *et al.*, 1996; van Busschbach *et al.*, 1997). If contributions of non-health factors are significant, the value function obtained may be invalid.

Table 2 summarizes the requirements for the analytical models required in step 1 and 2 in the two-step process to derive a value function. If results in step 1 are satisfactory, step 2 is different for naturalistic and multiattribute scenarios. For naturalistic scenarios, we can obtain a "best" estimate of each health-state scenario valued, but as these descriptions are at best at a nominal measurement level, for non-valued states no values can be obtained. In practice, interpolation techniques are therefore applied, based on the assumption that a non-valued health-state scenario can be safely put in between two other, successfully valued, states by a formal procedure (see for an example: Murray & Lopez, 1996).

With multiattribute scenarios, we may specify a value function, that is a quantitative prediction rule which for any possible health state of the classification (viz. any combination of the attributes) provides a value between 0.0 and 1.0. It follows that if major non-health effects are likely after step 1, and if such effects are desired to be reflected in the value function (e.g., age of the patient described in the scenario, the country of application of the rule), this may be enabled by the choice of a suitable mathematical function. Only if a multiattribute health-status classification system is relatively simple, a complete empirical valuation of all the optional health states can be constructed without a value function (Rosser & Kind, 1978; van Busschbach *et al.*, 1997).

Within this general approach of developing a value function many choices remain, for instance, on the exact form of the mathematical function. Except for one specific analytical model (see: next section), all other analytical models can estimate the value function (step 2), but are unable to support the determinant analysis of step 1.

Below we will discuss some of the most prevalent analytical models in order of appearance in Table 2, starting with the model that is suitable not only for step 2, but also for testing the data properties of step 1.

Functional measurement

Functional measurement originates from psychology and is based on a specific theory of human information processing (see: Appendix B). This theory assumes several, distinct, perceptual and judgmental processes during valuation procedures which can all be estimated, provided the data set satisfies the relatively large research design demands of this approach. For this purpose an appropriate design (experimental factorial design) and a related statistical technique (analysis of variance) are required (Anderson, 1970; Anderson, 1976). All types of stimulus data are allowed and also various functional relationships between stimulus and response (see: Table 3). The approach has the advantage of permitting the testing of the underlying subjective process by which respondents process information of multiattribute stimuli. It therefore allows for validation of the derived values. Unlike other methods, functional measurement analysis permits conclusions about the level of measurement (i.e., ordinal, interval, ratio) of assessments. For this approach the respondents have to value a *specific* (experimental design) set of health scenarios with varying levels of the attributes eventually combined with variations for each other (stimulus mode) factor of interest. The

Table 2
Four basic requirements for the analysis of valuation data at the Analytical Stage

Classification		Aspects
Class A: data properties	a1	Be flexible at the measurement level, preferably be able to test the *assumed* measurement levels or equidistancy, optionally after appropriate data transformation.
	a2	Be flexible over aspects such as the number of domains, the number of levels for the domains, valuation methods, response modes etc.
Class B: contribution of effects (step 1 only)	b1	Permit inference about the impact of health scenarios, more precisely, the separate stimuli (domains, levels, duration statements etc.) contained within the (multiattribute) scenario.
	b2	Permit inference about characteristic response behaviors of (subgroups of) subjects (e.g., risk attitude, time preference).
	b3	Permit inference about the separate impact of subjects' background characteristics (e.g., age, education, sex, health experience).
	b4	Permit inference about systematic effects of all other factors not related to scenarios and subjects; such as valuation procedure, stimulus mode and response mode.
Class C: functional relation of effects (step 2 only)	c1	Be able to estimate specific parameters for, at least, factor b1.
Class D: functionality	d1	Require a limited number of health scenarios and/or responses per scenario to be valued; this number mainly depends on the number of restrictions and other assumptions of the model.
	d2	Provide prediction rules in an easy way (parsimony criterium).
	d3	Require standard statistical software packages.

Table 3

Characteristics of the models at the Analytical Stage to analyze valuations of health outcomes elicited at the Valuation Stage (classification and their subdivision refer to Table 2)

Analytical Model	Class A: Data Properties		Class B: Determination of Effects (step 1 only)				Class C: Modeling of Effects (step 2 only)	Class D: Functionality			Analytical Software*	Specific Data Restrictions	Applicable Valuation Methods
	a1	a2	b1	b2	b3	b4	c1	d1	d2	d3			
Functional Measurement	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	1, 2, 3		all, except PC
Multiattribute Utility Theory	-	✓	-	-	-	-	✓	½	✓	✓	4		all, except PC
Multiple Classification Analysis	-	✓	✓	-	½	½	✓	½	✓	✓	1	no interactions	all, except PC
Multiple Regression Analysis	-	✓	½	-	½	½	✓	½	✓	✓	1, 2, 3		all, except PC
Paired Comparisons	-	-	-	-	-	-	.[a]	½	.[a]	-	1		PC
Conjoint Measurement	-	-	-	-	-	-	✓	½	✓	✓	1, 2	no interactions	PC
Unfolding Analysis	-	✓	-	½	-	-	.[a]	½	.[a]	✓	1		RS, VAS

✓ Aspect satisfied.
½ Aspect partially (under specific conditions) satisfied.
- Aspect not satisfied.
. Not relevant.
a Scaling of empirically valued scenarios only.

* 1 = SPSS, 2 = SAS, 3 = BMDP, 4 = special purpose program

method is equally suitable for step 1 and 2. The drawbacks of this technique are that a vast number of valuations have to be acquired and that thorough knowledge of the analysis of variance technique is essential.++ Summarizing the performance of functional measurement (for an overview of all methods, see: Table 3), we concluded that functional measurement is sufficiently flexible to detect the measurement levels, the functional relationships between the domains and to construct a valid prediction model based on these features (aspects a1, a2 of Table 2). From Table 3 it is also apparent that functional measurement is the only method which permits all four types of effect inferences, distinguished in Table 2 (b1 to b4). Admittedly, the simultaneous estimation of health scenario effect (b1) and even one of the additional three aspects (b2 to b4), requires substantial empirical data sets. Functional measurement is not particularly better to parametrize the effect of health domains (c1), but so far it is the only method which permits the testing of the underlying subjective processes by which respondents process information and to construct a prediction algorithm next to it. Apparent from the inevitable data demands and the need for specialized methodological and statistical expertise, its functionality is high (aspect d2 and d3).

From a theoretical point of view, we agree with Froberg and Kane (1989b): "(...) the functional measurement approach is clearly superior to the other designs".

Multiattribute utility model

This model is based on a general theory which has been developed in the field of (welfare) economics and is concerned with expressing the utilities of multiattribute entities (here: health outcomes) as a function of the utilities of each separate attribute. Like functional measurement the operational design and the selection of health scenarios follows from the theory. The stimulus should consist of multiple attributes (the economic equivalent of domains or dimensions), with ordered categories. Different functional relations are allowed. In this context the most popular function is the additive model without (main effects) or with interaction (usually only first-order) between the attributes. More saturated models were published (Fischer, 1979; Torrance *et al.*, 1982; Torrance *et al.*, 1996).

Usually these models assume independence among the attributes for parsimony. In health-status applications, the results do not in all situations support the additive model, and for these situations a multiplicative model or a model with interactions seems better (Feeny *et al.*, 1994) (see: Appendix C).

Multiattribute utility (MAU) theory (Keeney & Raiffa, 1976) offers a framework which is analytically founded on a combination of specific study designs, variants of regression analysis and formal tests to detect and model dependence among attributes (Fischer, 1979). In the usual MAU study design, the respondents are asked to evaluate each level of a particular attribute assuming all other attributes are held constant at some level (e.g., in EuroQol terminology: 22221, 22223; 22212, 22232; ...). Moreover, MAU models require a predefined set of (multi)attributes to be assessed to estimate the effects of the levels for all the attributes. However, under the assumption of dependency among attributes much more health states have to be valued. Especially for the extended models, the estimation of the parameters of

++ For empirical applications see: Veit *et al.*, 1982; Cadman & Goldsmith, 1986.

MAU models is demanding (although condition d3 is satisfied), representing the main limitation of these models. In addition, the assumptions made by the multiattribute utility theory are rather restrictive and no sound (say: independent) methodology is provided to validate the estimated weights (regression weights; see below) of the attributes of the value function, nor the constructed utilities (coefficient of determination; see below) of the health states themselves. This is not first and foremost a drawback of MAU theory, but just a limitation of the regression analysis by which all MAU models are estimated.

Some frequently-used instruments, for instance, the Health State Classification System (Torrance *et al.*, 1982), are constructed using this type of MAU model. Notwithstanding its limitations, MAU models have been widely used in the United States and in Canada (Keeney & Raiffa, 1976; Pliskin & Beck, 1976; Torrance, 1982; Torrance *et al.*, 1982; Froberg & Kane, 1989a; Feeny *et al.*, 1992).

Multiple classification analysis

This statistical technique (see: Chapter 6) in particular, seems suited to analyze data from multiattribute health-state scenarios. Multiple classification analysis (MCA) can best be seen as a combination of functional measurement and multiple regression analysis, although it is only capable of dealing with additive models. The algorithm of this technique is, as opposed to functional measurement and multiple regression analysis, based on an iterative estimation procedure. The technique is able to indicate the (in)appropriateness of the additivity assumption (see: Chapter 6) and is more flexible than standard multiple regression analysis. It offers meaningful estimates of the contributions of the separate attributes and their levels if the additive assumption holds. As opposed to the standard approach of estimating value functions like multiple regression analysis, multiple classification analysis treats the levels of a classification system standard as ordinal, rather than interval indicators. The selection of the health-state scenarios as well as the valuation task for the respondents, is identical to multiple regression analysis.

MCA is in particular convenient and appropriate for exploring multiattribute value functions that are thought to be additive. It offers for these types of value functions a clear overview of the properties and relationships between attributes and their levels.

Multiple regression analysis

This generally applied statistical technique is in particular suitable for survey data. The dependent variable is the assessment for the health scenarios, which is predicted by the independent variables (the attributes). The main difference with the multiattribute utility model is that multiple regression analysis poses no underlying theory and is solely based on the algorithm of the statistical technique. The valuation task for the respondents is identical to the two models discussed above, albeit that for the selection of scenarios, no specific demands are imposed. As before, only a carefully planned selection of valued health-state scenarios enables a valid inference of interaction terms between levels of attributes. A major disadvantage of the regression strategy is that it does not provide a way to determine the validity of the estimated regression weights. In particular, the expression by the model of the coefficient of determination (R^2) does not form an adequate test for the fit and validity of the prediction model because R^2 can be high even when deviations from model predictions are significant

and systematic (Anderson, 1971; Froberg & Kane, 1989a; Krabbe *et al.*, 1995). A limitation of most of the studies carried out using this statistical model is that these studies are not performed on the complete data but on aggregated (means, medians) health-state values, precluding any control for individual systematic differences (Torrance *et al.*, 1996; van Busschbach *et al.*, 1997) and other effects (the second approach described on page 39).

Others

Finally, there are a number of less-known strategies, that are all suitable, in theory, for multiattribute *and* naturalistic scenarios.

Paired comparisons — This is an indirect scaling technique originating from psychology (Torgerson, 1958; Kruskal & Wish, 1978; Lewis-Beck, 1995). An application of the paired comparisons method that focused on the valuation of health states was carried out by Hadorn *et al.* (1992). Another study showed that the paired comparisons method was not appropriate for multiattribute scenarios due to the specific features of these scenarios (Krabbe *et al.*, 1997).

Conjoint measurement — This is a scaling technique that stems from the area of marketing and consumer science (Luce & Tukey, 1964; Green & Rao, 1971; Louviere, 1988). Maas (Maas & Stalpers, 1992; Maas & Wakker, 1994) has used conjoint measurement within a clinical decision-making setting.

Unfolding — This is also an indirect scaling technique originating from psychology (Coombs, 1950; Coombs, 1964; Bossuyt, 1990; Lewis-Beck, 1995). The unfolding strategy, according to our knowledge, has never been used for constructing scales for health states.

The three model strategies above are not able to determine the effects of Table 2 (b1-b4). These scaling strategies are in effect indirect scaling (c1, Table 2) techniques, unsuitable for statistical inferences. Hakim *et al.* (1995) have recently introduced and compared new strategies in the field of health-state assessment, that are related to conjoint measurement. Multilevel models (Goldstein, 1995; Duncan *et al.*, 1998) may also offer a contribution to the analysis of elicited values to arrive at reliable value functions. The main characteristics of the statistical techniques discussed above are presented in Table 2.

V.2 — Analysis of the descriptive data

The goal of this operational procedure is to obtain a set of health scenarios in a way that adequately represents the health outcome of the problem under study. This procedure is different in naturalistic scenarios compared to multiattribute scenarios.

Regardless of the type of scenario used, in most clinical applications, only a small subset of actual health states is prevalent, hence needs to be valued later. For some clinical studies including heterogeneous patients at entry or follow-up, it may be a laborious task to summarize the descriptive data of patients, in particular if naturalistic scenarios are used.

While only a small subset of health states is prevalent in large population surveys (Kind *et al.*, 1998; Mayo *et al.*, 1998), a much greater number of states is prevalent in less healthy populations (Fryback *et al.*, 1993).

No formal methods are yet available to explore the representativeness of a set of naturalistic or multiattribute scenarios (with their frequency of occurrence) in a given context.

In our studies we relied on expert judgment, but reliable methods have to be developed if the QALY methodology is applied in decision making

V.3 — Estimation of the value for actual health states

The estimation of the value of all distinguished health scenarios (either naturalistic or multiattribute scenario) is straightforward if no additional non-health factors are taken into account. For naturalistic scenarios, we have nominal stimuli, hence no computations are required, only the assignment of descriptive data to one separate scenario for which the value was gathered previously. For multiattribute scenarios a computational rule (the value function) assigns a value.

So far, little attention has been paid to the confidence intervals of such computed values, which have to combine measurement error of the descriptions and estimation error of the value function (see for an example: Torrance *et al.* (1996). Usually the resulting values (of Q) are applied as error-free constants in the last analytical operation.

V.4 — Estimation of summary values (QALYs)

The next and final step in the decomposed QALY models consists of the determination of $U(Q, L)$, e.g., QALYs, or in the case of individual prognostic elements, the determination of $U(Q, L)^{N^*}$, e.g., QALEs (see, below). This final analytical step is only not necessary when $U(Q, L)$ or $U(Q, L)^{N^*}$ is directly determined, as is the case for the Single Profile QALY Model, respectively the Multiple Profile QALY Model.

The computation of summary values for health outcomes are already presented in this chapter. Eventually, adjustments such as age-weighting and discounting of life years can be performed as a concluding analytical operation (see: Chapter 10, p. 166).

Examples QALY Projects

In order to illustrate QALY models, we describe two projects in which we have participated.

Disability weights for diseases in the Netherlands

Recently an approach focusing on the quantification of the burden of disease and injury on human population has been developed by Murray & Lopez (1994; 1996). Their work (1997a; 1997b; 1997c; 1997d), which was first published in the World Bank's authoritative report "Investing in Health" (1993), resulted in a comprehensive set of estimates for the global and regional burden of disease resulting from premature mortality and disability. The quantified burden of disease was expressed as Disability-Adjusted Life Years (DALYs), lost due death, disease or injury (Table 1, Stage I). A replication of this study with the addition of some related methodological inquiries has been conducted in the Netherlands (Ruwaard & Kramers, 1997; Stouthard *et al.*, 1997).

In the Dutch study, DALYs were derived by a three-step procedure along the same lines as in Murray's study. First, a set of diseases ("indicator conditions") were described in a naturalistic manner (one year period), but in addition a disease label was added together with a EuroQol-based multiattribute description (Table 1, Stage II). This set of mixed scenarios was valued (the *Q* of that period) by a panel of experts (Table 1, Stage III). The person-tradeoff (PTO; this chapter, p. 36) method was applied (Table 1, Stage III) to obtain the weights for the indicator conditions. Second, to facilitate the estimation of disability weights for several hundred other conditions (described in a naturalistic way), a disability scale (e.g., visual analogue scale; Table 1, Stage III) was constructed with the valued indicator conditions as benchmarks. The remaining conditions were all interpolated on the disability scale by a second panel of experts (Table 1, Stage IV).

The final result of the DALY measure is a comprehensive list of disability weights for a large number of diseases. If combined with epidemiological data on the prevalence (*N*) of such conditions and their duration (*L*), such disability weights can be used to calculate Burden-of-Disease estimates (Table 1, Stage V). Murray's unit of measurement was called Disability-Adjusted Life Expectancy (DALE), which is a form of QALE (i.e., Partially Decomposed QALY Model; see below).* The use of this normative technique can be easily combined with specific interventions or other induced changes, as shown by Murray & Lopez (1996).

Treatment decisions for laryngeal cancer

Clinical decision making introduces an element into the general QALY model that is different in QALY modeling at the societal level. At the individual level, decision makers are always confronted with probabilities/uncertainties (*P*) of treatment outcomes. So, the element *uncertainty* is inherently introduced in the QALY model (ex ante modeling).

The first step towards clinical decision analysis involves identifying the treatment options and the potential outcomes of each. This information can be structured, usually in the form of a decision tree (Figure 12). Such a tree consists of nodes, which describes choices, chances and outcomes. Choices represent the strategies available to the physician and the patient. The probability that each outcome will occur (*P*) if a particular strategy is employed is expressed by probabilities (chances) for each such node. The relative value of each treatment outcome (*Q*) has ultimately to be supplied by the patient him/herself. Finally, analytical operations are performed for the three key elements of the general QALY concept (i.e., health status, life years and probability) to arrive at Quality-Adjusted Life Expectancies (QALEs), as a summary measure to express the individual preference for a treatment. Again this measure is a QALY-analog.

In our study, we have compared this analytical approach based on an additive elaboration of the Single Profile QALY Model (Chapter 8). Furthermore, we have studied the effect of (additional) clinical information on the preferences of the treatment outcomes (Chapter 7) and the stability of such preferences by replication of the valuation tasks (Chapter 9).

---

* More precisely: DALYs are QALYs lost, compared to some standard (e.g., expected life years), where the decomposed model is adopted.

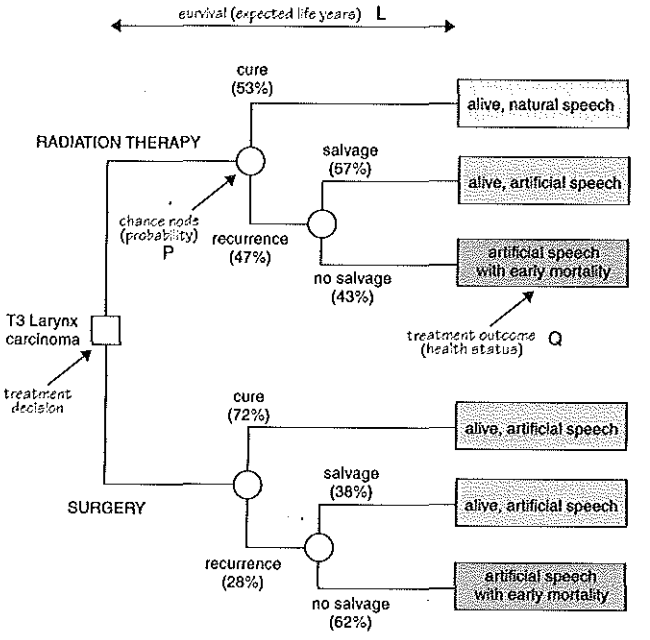


Figure 12
Decision tree for a 65-year-old man with T3 laryngeal cancer; 2 treatments, 3 different treatment outcomes, 6 different trajectories (health profiles)

PART I

Chapter 3

On the Equivalence of Collectively and Individually Collected Responses

Standard-Gamble and Time-Tradeoff Judgments of Health States

Paul F.M. Krabbe, Marie-Louise Essink-Bot, Gouke J. Bonsel
Medical Decision Making; 16:120-132, 1996

Reprinted with permission from Medical Decision Making

Abstract

The standard-gamble and time-tradeoff methods for valuing health states were compared in a multifactorial design with 104 student volunteers. The main aim of the experiment was to compare average individual responses with group responses for the same tradeoff tasks. Group responses were collected using an interactive voting system. The standard EuroQol system was used to describe the health states to be valued. Generalizability theory was used to analyze the results.

The averages and median values of the individual responses differed from the interactively collected group values only for the more severe health states. The results showed almost identical results for the two methods, but the time-tradeoff was found to be more consistent than the standard-gamble.

The authors conclude that 1) there is significant similarity between the results of individual and collective response modes, and 2) the standard-gamble and time-tradeoff methods produce almost equivalent values, despite their different conceptual backgrounds. In this study the aggregated individual responses and the collective response proved to be sufficiently similar to support the validity of using aggregated individual valuations as a measure of the valuation of the group.

Key words: collective judgment, standard-gamble, time-tradeoff, methodology, voting system, EuroQol, generalizability theory

Introduction

Health status has become one of the important outcome measures in the evaluation of medical interventions, in addition to the more traditionally used mortality and symptom-oriented measures. Generic (i.e., comprehensive and non-disease specific) health-status measures commonly take the form of profiles, e.g., the Nottingham Health Profile (NHP) and the MOS Short-Form 36 (SF-36). However, application of health-status data in medical evaluation research, in particular in economic analysis, requires us to go one step beyond mere profile descriptions of health status. A descriptive measure with a physical dimension (A) comprising 3 levels (1 = optimal, 2 = intermediate, 3 = worse) and a psychosocial dimension (B) with three analogous levels may, for example, generate a profile of A_2B_2 for patient X, and A_1B_3 for patient Y. However, if we are to judge whether patient X is better or worse off than patient Y, and if so, by how much, a single summary measure is required. Such summary scores may be obtained by using a valuation procedure for each health state. The resulting values may be used to combine life years with quality, both in economic evaluations [e.g., in heart transplantation (van Hout *et al.*, 1993a), coronary artery bypass grafting (Williams, 1985), administration of ACE-inhibitors (van Hout *et al.*, 1993b)] and in assessing the burdens of illness upon populations in public health modeling [e.g., the World Bank report "Investing in Health" (World Bank, 1993)].

In an empirical procedure designed to obtain values for health states, subjects were requested to indicate how good or how bad certain health states were for them, and to rank these health states according to their degrees of undesirability by means of methods such as the standard-gamble (SG) and time-tradeoff (TTO). A major issue within the scientific field of health-status valuation is, first, whether individual value responses can be aggregated to reflect group values, and if so, how (Hilden, 1985; Loomes & McKenzie, 1989). Second, there is still considerable debate as to which method, i.e., SG or TTO, is the most valid and precise.

In this paper we compare SG and TTO data collected as group responses and individual responses. We address the following research questions:

- *Are aggregated individual value responses comparable to collective response of the group as a whole?* We hypothesized group response to be similar to the median of the individual responses.
- *How do SG and TTO compare in terms of equivalence (validity)?*
- *What are the sources of measurement error for these two elicitation methods?*

To answer these questions, we conducted an experimental study with a multifactorial design (see below).

We used the EuroQol descriptive system for health status, as our study was part of the program of the EuroQol Group.

In 1987, the international and multidisciplinary EuroQol Group (EuroQol Group 1990; Essink-Bot *et al.*, 1993) was established. Since that time the Group has developed a short generic instrument to describe a patient's health status based on five dimensions, each comprising three levels. Health states were generated by combining discrete levels from each

dimension. Valuations for sets of EuroQol health states were elicited from population samples to enable the intended use of EuroQol in economic evaluation. The current EuroQol valuation questionnaire appeared to be suitable for postal surveys without interviewer support and produced consistent values on a selected set of health states by means of a visual analogue scale (VAS). Population surveys in the United Kingdom (EuroQol Group, 1990) The Netherlands (Essink-Bot *et al.*, 1990), Sweden (Brooks *et al.*, 1991) and Norway (Nord, 1991) showed international similarities of valuations and characteristics of the responses. The sensitivity of the valuations for sociodemographic variables appeared to be low (Essink-Bot *et al.*, 1993), and little bias may be expected from the non-responders (Essink-Bot *et al.*, 1993). Contextual effects were minimal and reproducibility was satisfactory (van Agt *et al.*, 1994).

Material and Methods

Organization

Extensive pilot studies preceded the experiment, which included two sessions, separated by a ten-day interval. Students were recruited by handouts. The same group of 104 students participated in both sessions. For full participation they were paid approximately $ 65 (1993). Both sessions consisted of a sequence of predominantly valuation tasks (the results of the experiments presented here are part of a more elaborate experiment). The valuation tasks were deliberately interspaced with unrelated questionnaires, for example, on the moral acceptability of genetic manipulation, to avoid weariness and irritation due to monotony.

All participants were seated in a lecture hall with due space between them. The different methods were preceded by a similar verbal explanation of the method and a few test judgments. The stimuli (the health state to be valued and, with the collective tradeoff tasks, the alternative for comparison) were always presented by slide projection. During the presentation, the instructors (GJB, MLE-B) repeated the nature of the particular tradeoff task with each stimulus, to avoid any blurring of the concepts of SG and TTO. Responses were recorded by pencil and paper for the individual tasks, and by means of an electronic interactive voting system for the collective tasks.

Conventional aggregation of the valuations of all individuals for one given health state by computation of the mean or the median is indicated by IND. This standard method was compared to a method which aimed at a genuine group or collective response by means of an interactive voting system. The collective response mode is indicated by COL.

Health-state descriptions

The EuroQol concept of health status consists of five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each dimension has three levels, "no problems" (1), "some problems" (2) and "severe problems" (3). Health-state scenarios are produced by the selection of one level for each dimension (the best health state is thus represented by 11111). Theoretically, this set of dimensions and levels of the EuroQol

instrument allows for 243 (3^5) different health-state descriptions. The EuroQol Group selected 13 of these "scenarios" as a standard set as a basis for experiments. We selected 11 states from this standard set (excluding: 11111 and 33333). Previous results had indicated, however, that this set did not evenly cover the continuum between 0 and 100 (Essink-Bot *et al.*, 1990; van Busschbach *et al.*, 1994). Two "gaps" existed in the value range of this standard set of states, so we therefore added two other health states (12212 and 33332).

Standard-gamble

The SG concept is derived from the von Neuman-Morgenstern (vNM) utility gamble (von Neumann & Morgenstern, 1953). The method (Torrance *et al.*, 1972) is essentially an iterative paired comparison. The participant is presented with two alternatives, and asked to select the preferred one. One alternative offers the participant a specified certain outcome, while the other alternative offers a gamble with mutually exclusive probabilities for two reference outcomes. Conventionally, SG is operationalized as a choice between A, the certainty of being stationary in a specific lifelong impaired health state (the state to be valued), or B, the uncertain result of an intervention, for example a surgical procedure, with two reference outcomes. These are: a probability (p) of instantaneous and lasting improvement to perfect health, or a probability ($1-p$) that the operation will fail, resulting in immediate death. By varying the p-level of the uncertainty outcome, the point of indifference between the two alternatives A and B is determined. By combining the probability values with the utility values of the reference outcomes, the utility of the stationary state is established. The method fails if the state to be valued falls out of the range covered by the two reference outcomes (for example, if in the conventional operationalization a state is valued as being worse than death). To overcome such a situation, an adaptation of the presentation of A and B is necessary.

In the present experiment, two slides were shown simultaneously to generate a collective response using SG (SG_{col}). The certainty of the stationary chronic health state to be valued was shown on the left, while the gamble was shown on the right. The gambling probabilities of the alternative option were varied in steps of at least 2% upwards or downwards, depending on whether the gamble or the certainty was preferred, until the participants no longer preferred one over the other. Each individual expressed his or her preference for one or another alternative by means of the voting system. The first alternative option for SG was the same for all health states to be valued, namely a 50% chance of being in the "best imaginable health state" (described as "perfect health" in other studies) and a 50% chance of being in the "worst imaginable health state".* Figure 1 shows an example of the SG_{col} presentation. The state to be valued is shown on the left (in this example EuroQol state 21232). The alternative option on the right shows a gamble with an 88% chance of the "best imaginable health state" and a 12% (100%–88%) chance of being in the "worst imaginable health state". Is was clearly stressed to the participants that both outcomes arising from the gamble would involve chronic health states.

* Instead of the convention of using being dead as the bad reference outcome, we used "worst imaginable health state", for reasons that are explained later. The consequences of this strategy are also discussed.

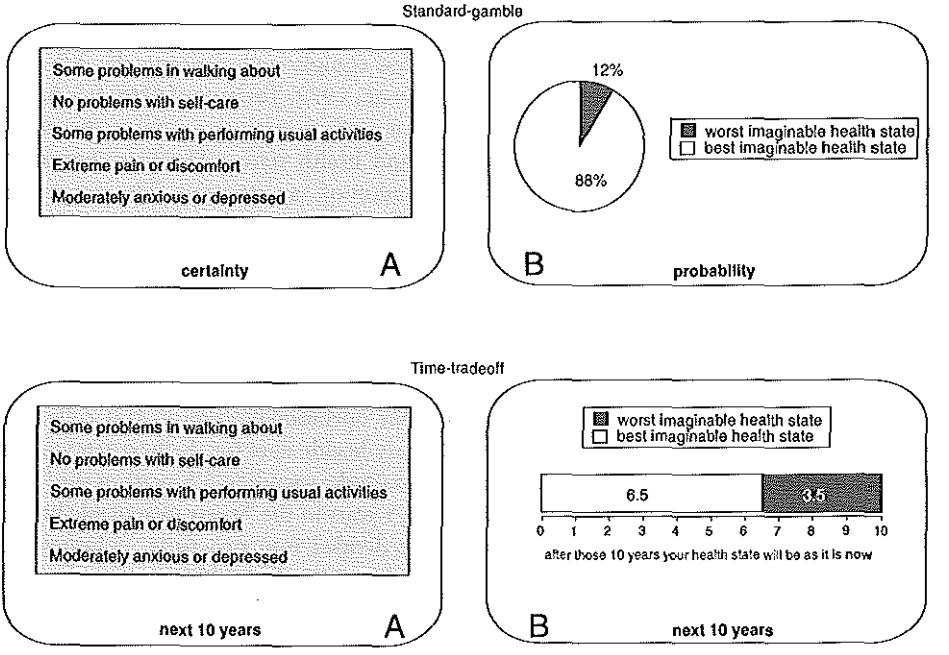


Figure 1
Examples of the slide presentation of the stationary health states to be valued (left) and the alternative options

The description of the SG measurement process includes four specifications: the duration of the state, the exact prognosis following the state, the age of onset for the state, and whether or not the state applies to the subject himself/herself or to someone else. In addition, the subject's economic well-being should not confound the measurements (Torrance, 1986). All the requirements were fulfilled in this study. Due to the fact that all participants were students, the age of onset was similar, i.e., approximately 25 years. For the SG method all health states were chronic, so duration and prognosis were fixed and were therefore the same for all participants. In The Netherlands all citizens have free assess to standard medical care. Consequently there is no direct confounding with economic well-being (purchasing power for medical care is homogeneous).

For the elicitation of individual responses (SG_{ind}), the state to be valued was presented on a slide. Each individual responded conventionally by dividing a "probability pie" into two complementary parts using pencil and paper. Thus the individual presentation of the probability pies corresponded exactly to the slide presentation in the collective session.

Time-tradeoff

The method TTO was developed by Torrance as a less complicated, conceptually different although equally sound, alternative to SG (Torrance *et al.*, 1972; Torrance, 1976). It is based on tradeoffs similar to those of the SG, but the concept of uncertainty is omitted from the TTO method. The participant trades off survival and health status. The first alternative specifies a (suboptimal) health state with a given duration of, say, ten years. The competing alternative offers a better health status (conventionally optimal health) of shorter duration, conventionally followed by death. The point of indifference is reached by varying the duration spent in perfect health. By combining the duration value and the values of the reference outcomes, the utility of the stationary state is established.

The operationalization of TTO in the present experiment involves the following alternatives. The first option was five years in the "best imaginable health state" followed by five years in the "worst imaginable health state". Our operationalization of this collective response (TTO_{col}) is shown in Figure 1. The stationary health state for the next ten years is shown on the left. The alternative in this particular example, spending 6.5 years in the "best imaginable health state" followed by the remaining 3.5 years in the "worst imaginable health state", is shown on the right. A bar, proportionally divided into two parts, indicates the numerical presentation of the number of years. For both options the health state would return to its present form after 10 years. The years in the alternative option were varied in steps of 0.5 years.

Individual responses (TTO_{ind}), were collected from participants using pencil and paper. Each individual participant was asked to divide a "duration bar" into two parts. The collective response were obtained by showing two slides simultaneously, i.e., both the stationary health state to be valued and the specified alternative. The same procedure as that used in the SG method was used to achieve the point of indifference.

Within each separate SG and TTO experiment, the states to be valued were presented in a randomized order to avoid memory effects.

The interactive voting system

A voting system with a button-box at each seat was installed in the lecture hall. The system allowed for a dichotomous response (for example, preference for either alternative A or alternative B), and for a numerical response (1 to 10, in this study used only for identification control). The number of participants and the percentages of participants voting for the stationary state to be valued A and for the method-specific alternative option B was continuously displayed on a monitor which was visible to the investigators only. The first slide with the alternative options from which all participants had to choose individually was the 0.5/0.5 option for SG and the 5/5 option (years) for TTO. Option B was varied until the indifference point between slide A and B was achieved. Indifference was defined as a situation in which the number of votes for one option was between 48% and 52%. The next option B was based on the magnitude of preference for a particular alternative, and, for all subsequent bids, on the options already offered. If, for example, during the first voting cycle, 72% of the participants voted for A and 28% for B for a particular EuroQol health-state description, the

alternative option B was varied into a more desirable outcome during the next voting cycle. This involved reducing the "worst imaginable health state" percentage for SG and reducing the number of years to be spent in the "worst imaginable health state" for the TTO. Based on the outcomes of the previous assessment, a simple software program selected the next slide for alternative option B that would lead most efficiently to equivalence between options A and B. For example, the succession of presentations of the option B for health state 12212 was, for the SG method, 0.5/0.5 (13% preferred B), 0.12/0.88 (58%), 0.32/0.68 (28%), 0.22/0.78 (40%), 0.16/0.84 (51%). After five bids, indifference (49% A, 51% B) was reached for this health state, and, with the last proportions, 0.16 vs. 0.84, the collective valuation was gathered to be 84. In the experiment, the collective response stabilized after four to six bids. We regarded the score at the indifference point as a genuine point estimator of the group response.

To investigate the test-retest reliability of the collective response procedure, the SG_{col} session was repeated during the second session.

The position of "dead" in SG and TTO

In both the SG and the TTO, the state "dead" occupies a specific position. In the SG the calibrating gamble is usually between perfect health and "dead". In Torrance"s original operationalization of TTO, "dead" follows the shorter period in perfect health. In the present experiment, "dead" was replaced by the "worst imaginable health state". Similarly, for the TTO, the period in "perfect health" was followed by a complementary period in the "worst imaginable health state". These two periods together were equal to the total duration of the stationary period.

The main reason for selecting "worst imaginable health state" relates to the primary objective of our study, i.e., the comparison of the utility elicitations in a collective response mode and an individual response mode. Normally, when a health state is valued as being worse than "dead" (indicated by a preference to die immediately instead of living any number of years in the state to be valued), a modification of the SG and TTO methods is necessary (Patrick *et al.*, 1994). In the present study, this should be the replacement of the "worst imaginable health state" with a description of the EuroQol health state valued worse than being dead and the replacement of the stationary health state (normally the state to be valued) with "dead". However, the collective response mode precluded the use of such a complex "mirror" procedure for states worse than "dead".

A further reason for selecting "worst imaginable health state" was that in the conventional operationalization of SG and TTO, "dead" only serves as a benchmark. "Dead" is not an essential part of either method; neither is the use of perfect health at the other extreme (Torrance, 1986). Logically any two pairs of reference states are suitable so long as they "embrace" the state to be valued and their utility values are known. As with the SG, the TTO procedure should allow for the use of reference states other than those used conventionally. (This needs the imputation of other utility values in the final calculations; see, Llewellyn-Thomas *et al.* (1982) for a test of this assumption in the SG). If for SG and TTO reference states other than perfect health and/or dead are used, the utilities obtained with such non-standard operationalizations are different and need rescaling factors to be comparable with utilities obtained with the conventional SG and the standard TTO (see: Appendix 3.1).

Furthermore, it is a matter of preference or convention to anchor the value of "dead" at 0 (zero) (Weinstein *et al.*, 1980). Inevitably this convention leads to the assigning of negative values for the worst health states, regardless of the health-description system used. In QALY calculations, negative values of health states may result in complicated computations.

Finally, we wanted to be able to compare SG and TTO data with standard EuroQol VAS data (trade-off techniques vs. a rating scale technique). In the EuroQol standard questionnaire "death" is rated through a separate valuation task. With the additional measurement of the value for "dead", scores on the "healthy-worst imaginable health state" scale can be transformed to a 0 – 1 perfectly healthy-dead scale of values.

In order to use the results of this study, for example, for the computation of QALYs, the utility for the "worst imaginable health state" would have to be determined (Llewellyn-Thomas *et al.*, 1982; Torrance, 1986). This drawback does not preclude us from investigating the main aim of this study, namely the degree of comparability between the two methods and the two response modes, by statistical and psychometric methods.

Statistical analysis

Means and medians were computed for the individual responses. Paired t-tests between SG_{ind} and TTO_{ind} for all health states were carried out and effect sizes (Cohen, 1977) were computed.

Pearson's product-moment correlation coefficient (ρ) and Spearman's rank-correlation coefficient (ρ_r) are the parameters most frequently used in analyzing parallel (equivalence) data. The first is suitable for interval or ratio data, as the ρ_r coefficient is more appropriate for ordinal data or data of a higher measurement level, which do not satisfy distribution requirements for ρ. Despite their popularity, both are for obvious reasons essentially insufficient for testing equivalence of single responses (Deyo *et al.*, 1991). Nevertheless, these statistics are presented here to allow for comparison with other studies.

In order to test equivalence of single valuations, the intraclass correlation coefficient (ICC) (Deyo *et al.*, 1991) is considered to be more appropriate.† The ICC takes into account the variability due to two systematic sources of bias, i.e., a level effect between two measures and a linear transformation similar to ρ. Furthermore, the ICC is more flexible compared to the other two correlation coefficients as it may be estimated in designs with multiple retests or with more than two raters.

Conventionally, the equivalence of valuation methods has been investigated by plotting the population's mean or median health-state values for each separate method and by subsequent regression analysis (see, e.g.,: Torrance *et al.*, 1972; Torrance, 1976). Although a considerable amount of information is lost by using aggregated data for regression analysis, we

† Our use of the correlation coefficients (ρ, ρ_r, ICC) for estimating coefficients of equivalence can be seen as measuring the criterion validity. In that case SG would be treated as the criterion ("gold standard") and TTO as the measuring instrument that corresponds to the criterion. By definition, the criterion must be a superior, more accurate measure of the phenomenon if it is to serve as a verifying norm. In this paper SG is not seen as a superior method for measuring preferences of health states but as a competitive method to TTO. For this reason, we have treated the two methods as equivalent, inducing us to speak of convergent validity.

present the results of this approach for comparative reasons and to study the relationship between the different methods and response modes.

An extension of the concept underlying the ICC is Generalizability Theory (G-theory) (Crocker & Algina, 1986; Shavelson *et al.*, 1989; Streiner & Norman, 1995). Where ICCs deal with two-way designs (subjects × raters or occasions), which makes the ICC a special case of G-theory, G-theory deals with *n*-way designs and provides an even more flexible, practical framework for examining different sources of measurement error. G-theory extends classic test theory by recognizing and estimating the magnitudes of the multiple sources (facets in G-theory language) of measurement error. In the present study, the relative contributions to "health states" by the facets "methods" and "participants" and the one- and two-order interaction terms of these two facets with "health states" are estimated within G-theory and named "variance components". From these estimated variance components, a generalizility coefficient $\hat{\rho}^2$, analogous to the standard reliability coefficient, can be calculated for SG and TTO (see: Appendix 3.2).

Both ICC and the G-theory can be implemented within the analysis of variance (ANOVA) framework. Deyo *et al.* (1991) described a simplified computation method for estimating the ICC, which can be managed even with a pocket calculator. G-theory studies are yet not available as a module of conventional statistical software packages, but they require a special-purpose program or simple adaptation of existing ANOVA modules. In our study we derived the necessary estimations of variance components from 8V of the BMDP-software package (BMDP, 1992).

Results

Response

Of the 104 participants, 46% were male. The mean age was 22 (SD = 2.48) years. All were students, 71% were medical students. SG_{ind} and TTO_{ind} each took approximately 15 minutes to complete. The complete series of bids for all 13 health states for the collective experiments, until collective indifference as defined in section "the interactive voting system" was reached, took approximately 40 minutes each. The feasibility of these experiments was satisfactory, although at the end some participants complained of weariness. Judging from participants' remarks and from the absence of learning effects, we regarded memory effects to be highly unlikely.

Descriptive statistics

Table 1 shows the results of the experiments after linear transformation to a uniform 0-to-100 scale (TTO_{ind} = 10 × score; TTO_{col} = 10 × score; SG_{ind} = 100 × score in degrees/360). Medians of SG_{ind} and TTO_{ind} are presented for appropriate comparison with the collective method. The order of presentation of the 13 health states in Table 1 is arbitrarily based on the SG_{ind} values. There is a noticeable similarity between the means and the medians of the IND

values. Except for one health state, the means of the health states were statistically significantly different for SG_{ind} and TTO_{ind}. However, the effect size indicated that for most of the health states those differences were small.

Generally, the individually and collectively aggregated values appear to have the same structure. The similarity of SG and TTO values is observed for both response modes. Differences between individually and collectively aggregated values are more obvious. In particularly, the health states 21232, 22323, 22233, and 33332, are valued differently to some extent within the two response modes. On average, the individual response to these worse states is higher than the collective response. The orders of the 13 health states for the two response modes shows approximately similar results for the SG and the TTO.

Table 1

Descriptive statistics on a 0 – 100 scale for the two elicitation methods standard-gamble (SG) and time-tradeoff (TTO) (individual and collective response modes) and paired t-test (p-values) and effect sizes (ES) for mean values of SG and TTO (individual response mode)*

	Individual Responses (session 2)					Collective Response		
	Mean			Median				
Health State**	SG	ES***	TTO	SG	TTO	SG (session 1)	SG (session 2)	TTO (session 1)
12111 (1)	96.2	0.23	94.5	97.8	97.0	96	96	95
11211 (2)	95.5	0.48	92.5	97.2	95.0	94	96	90
21111 (3)	94.5	0.22	92.8	96.4	95.5	94	96	90
11112 (4)	93.8	0.02	93.6	97.2	96.0	98	96	90
11121 (5)	93.3	0.17	91.8	95.8	95.0	96	96	90
11122 (6)	88.8	0.23	86.0	92.1	87.5	90	88	85
12212 (7)	81.9	0.22	78.7	86.4	80.5	84	88	80
32211 (8)	79.2	0.33	73.1	85.0	80.0	70	68	70
21232 (9)	65.2	0.28	59.2	66.5	60.0	44	58	45
22323 (10)	64.6	0.16	60.9	66.1	60.5	44	50	40
33321 (11)	53.7	0.23	47.9	54.2	46.5	44	50	50
22233 (12)	53.7	0.25	44.9	50.0	41.5	20	38	25
33332 (13)	34.5	0.27	28.0	33.1	20.0	6	12	5

* All differences were significant (t-tests: p-values < 0.05) except that for health state "11112".

** The EuroQol concept of health status consists of five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has three levels, "no problem" (1), "some problems" (2), and "severe problems" (3). Health-state scenarios are produced by the selection of one level for each dimensions (the best health state is thus represented by 11111).

Measures of equivalence

Three different correlation coefficients were computed as indices for the convergent validity between the SG and the TTO for all 13 health states, based on the individual valuations of the 104 participants. The overall means of the correlation coefficients for all health states show minor differences for the three correlation methods (Table 2). The overall ICC is slightly lower than Pearson's due to a small but distinct method level effect. The test-retest ICC for the SG_{col} was 0.97, indicating excellent reproducibility of the group responses. Figure 2 shows the results of regression analyses between collective responses and individual responses and between the SG and the TTO. Equations for the regressions of Figure 2 and for some regressions not depicted in Figure 2 are presented in Table 3. Regressions between collective responses and individual responses are presented for the median values only (Figure 2a). Coefficients of determination (R^2) are high, 0.97 and 0.96, respectively, for the regressions $SG_{col} - SG_{ind}$ and $TTO_{col} - TTO_{ind}$. The worse health states are valued somewhat lower in the collective sessions.

The scatter gram of method-specific median values for the individual health states (Figure 2b) shows that all SG_{ind} and TTO_{ind} median values are on the regression line. SG_{ind} is equal to TTO_{ind}, except for a small linear transformation (0.86), indicating that TTO_{ind} valuations are a little lower than SG_{ind} valuations (method effect). The coefficient of

Table 2

Convergent validity: the amounts of equivalence between standard-gamble (individual response mode) and time-tradeoff (individual response mode) as measured by intraclass correlation coefficients (ICC), Spearman rank-correlation coefficients ρ_r and Pearson product-moment coefficients ρ for 13 health states based on 104 respondents' valuations

Health State*	ICC	(ρ_r)	(ρ)
12111 (1)	0.55	0.69	0.63
11211 (2)	0.52	0.64	0.66
21111 (3)	0.69	0.69	0.73
11112 (4)	0.55	0.64	0.60
11121 (5)	0.80	0.76	0.82
11122 (6)	0.75	0.74	0.79
12212 (7)	0.51	0.61	0.52
32211 (8)	0.63	0.65	0.66
21232 (9)	0.65	0.68	0.68
22323 (10)	0.74	0.71	0.75
33321 (11)	0.70	0.72	0.72
22233 (12)	0.69	0.70	0.71
33332 (13)	0.65	0.69	0.68
Mean, all states	0.65	0.67	0.69

* The EuroQol concept of health status consists of five dimensions: mobility, self-care, usual activities, pain/discomfort, and mood. Each dimension has three levels, "no problem" (1), "some problems" (2), and "severe problems" (3). Health-state scenarios are produced by the selection of one level for each dimensions (the best health state is thus represented by 11111).

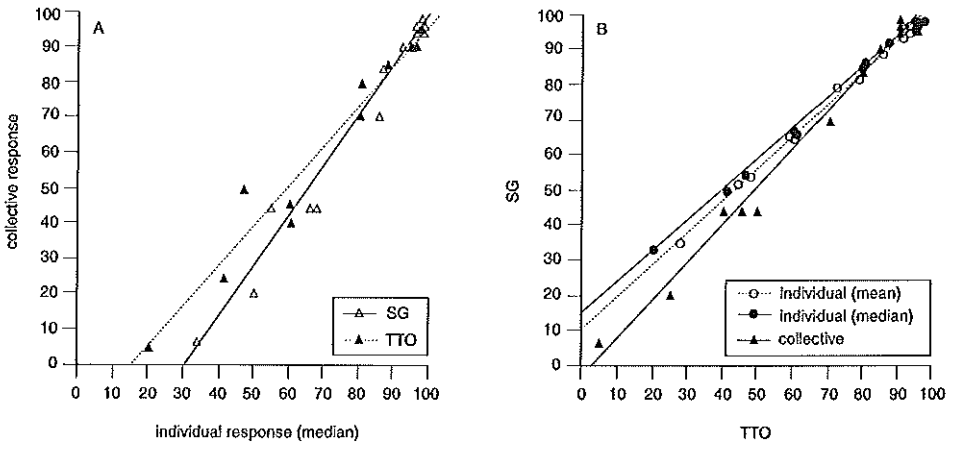


Figure 2
Scatter gram with regression lines of 13 health states; A: individual-median response versus collective response (standard-gamble and time-tradeoff method), B: time-tradeoff versus standard-gamble (individual-means, individual-medians and collective response)

Table 3
Equations of the regressions (individual response mode vs. collective response mode, standard-gamble vs. time-tradeoff) on mean and median valuations for the 13 health states

Comparison	Prediction Based on Means	R^2	Prediction Based on Medians	R^2
Individual vs. collective				
Standard-Gamble	$SG_{col} = -50.6 + 1.55 \times SG_{ind}$	0.97	$SG_{col} = -45.8 + 1.45 \times SG_{ind}$	0.97
Time-Tradeoff	$TTO_{col} = -28.0 + 1.29 \times TTO_{ind}$	0.95	$TTO_{col} = -18.0 + 1.14 \times TTO_{ind}$	0.96
Standard-Gamble vs. time-Tradeoff				
Individual	$SG_{ind} = 10.0 + 0.91 \times TTO_{ind}$	1.00	$SG_{ind} = 15.0 + 0.86 \times TTO_{ind}$	1.00
Collective	*		$SG_{col} = -3.54 + 1.08 \times TTO_{col}$	0.99

* The collective method essentially produces a median.

determination is near 1.0. Regression between TTO_{ind} and SG_{ind} performed with mean values yields the same outcome. Including the regression of collective responses, R^2 is 0.99, despite identical values of three health states for the SG_{col} method, whereas TTO_{col} showed minor differences.

Sources of measurement error

We performed G-theory analyses with the object of measurement (health states) and the two facets (participants and methods). G-theory allowed us simultaneous estimations of the effects of these facets on the 13 health states and we assumed that for each separate health state the 104 participants would produce equivalent values with both methods, SG_{ind} and TTO_{ind}. Deviations from this assumption were seen as measurement errors. Results of this analysis, based on the seven sources of variance (H, P, M, H × P, H × M, P × M, residual), are shown in Table 4. Starting with the primary effects (H, P, M), 56% of all variance was explained by the 13 health states (H). The contribution to measurement error of the participants (P) was relatively small; only 9% of variance was attributed to a systematically different valuation of the participants for all 13 health states. The systematic difference between the methods SG_{ind} and TTO_{ind} was negligible, i.e., M = 1%. The three first-order interaction terms (H × P, H × M, P × M) were responsible for 25% of measurement error. Interaction term H × M accounted for 0% of the variance. This finding supported the preceding regression analyses, which showed that the aggregated values for SG_{ind} and TTO_{ind} had a perfect linear relationship with each other in this study. The largest term of measurement error was H × P (22%). This means that the major part of the measurement error resulted from *some* participant's valuing *specific* health states differently compared with other participants and was *irrespective* of the method used. Non systematic error and the H × P × M second-order interaction (some participants valued a specific health state differently for one of the methods)

Table 4

Generalizability Study: sources of measurement error for the valuations of the 13 health states by 104 participants by the two elicitation methods, standard-gamble and time-tradeoff (individual response mode)

Sources of Variation	Sum of Squares (ss)	Degrees of Freedom (df)	Mean Squares (ss/df)	Estimated Variance Component	%
Health states (H)	1,141,064	12	95,088	455.38	56
Participants (P)	234,150	103	2,273	73.25	9
Methods (M)	9,550	1	9,550	6.66	1
HP	455,879	1,236	368	184.42	22
HM	3,041	12	253	1.74	0
PM	37,869	103	367	22.73	3
Residual (HPM, ε)	89,232	1,236	72	72.19	9

Table 5
Variance components (individual response mode) for the two elicitation methods, standard-gamble and time-tradeoff, and generalizability coefficients (individual response mode); measuring the internal consistency of the valuations of the 13 health states

	Standard-Gamble		Time-Tradeoff	
Sources of Variation	Estimated variance Component	%	Estimated variance Component	%
Health states (H)	415.41	56	497.10	64
Participants (P)	90.45	12	78.77	10
Residual (HP, ε)	238.80	32	202.23	26
Generalizability coefficient ($\hat{\rho}^2$)				
Per participant	0.56		0.64	
For the group (n = 104)	0.99		1.00	

were subsumed together within a small residual term (9%). A visual impression of the contributions of variance of the three sources of variance and their interactions is provided by the Venn diagram in Figure 3.

Generalizability theory also allows for a closer look at the reliabilities of the two valuation methods (Table 5 and appendix B). Separate estimation of the generalizability coefficient yielded an internal consistency coefficient of 0.99 for the SG_{ind} health-state valuations (based on 104 participants). This internal consistency coefficient, a specific type of reliability, stands for the precision of the valuations based on the aggregated individual responses. For TTO_{ind} the internal consistency coefficient of the health-state valuations based on the group level was 1.0.

Additionally, one of the advantages of G-theory is that it allows us to estimate internal consistency coefficients based on individual level. We have estimated these coefficients, which proved, of course, to be much lower for both methods (SG_{ind} = 0.56; TTO_{ind} = 0.64).

Conclusions and Discussion

The first research question addressed the comparison of two response modes. The collective response was assumed to produce a value that was comparable to the median of the individual outcomes. The specific nature of this group response (one point estimator, no dispersion by definition) precluded statistical testing of the difference between the two response modes and also limited the use of other explanatory statistical techniques. Generally, median values of individual responses were similar to the group values, with the bad health states as exceptions. We considered the following explanations for the somewhat different response for the bad health states.

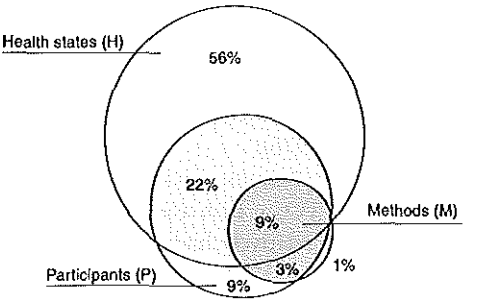


Figure 3
Venn diagram: contributions of the different sources of variance (object of measurement H and the two sources of measurement error P and M); 56% variance uniquely attributable to the valuation of the 13 health states, 9% of variance uniquely attributable to idiosyncratic responses of the 104 participants to the health states, 1% of variance uniquely attributable to a systematic difference of valuation between the two methods standard-gamble and time-tradeoff; all other percentages are combinations of the three unique sources of variance

Range compression. Each collective valuation experiment started with a 0.5/0.5 alternative, i.e., 50% good health versus 50% bad health. Thus, the reference was different compared with the individual experiments, where a blank "probability pie" or "ten-year duration bar" was presented on paper. Under the conditions of the individual valuation experiments, the participants were not confronted with the 0.5/0.5 situation and may consequently have been more cautious, giving low values to health states, with values close to 50 and lower.

Response shift. In the individual response mode, bad health states were valued with considerable individual variability. The cyclic process of the collective mode may influence the consistency of individual choice. If most participants prefer the alternative state when making the first comparison, the next comparison will be between the health state to be valued and a more severe alternative state than previously. Participants who, when making the first comparison, prefer the health state to be valued, may be aware of the discrepancy between their own valuations and those of the majority of the group, and shift towards the majority judgment in the second cycle. However, the number of instances of this type of inconsistency for the 13 health states were computed for all individuals for each health state and method. Per health state, these values ranged from 0% – 7% for both SG_{col} and TTO_{col}. Thus, response shift can hardly have been responsible for the differences between the two response modes.

Forced consideration. In the individual response mode, we assumed participants would take their own health status as a latent reference due to the fact that this is a pencil and paper task, whereas the collective response mode more forcefully encourages participants to examine only the health state to be valued (slide presentation) and the selected alternative for several cycles. Hence, the cognitive processes underlying the two response modes are probably different, which more generally may explain the common difference between tradeoff techniques and VAS.

Addressing the second research question, we compared the SG and the TTO and examined the equivalence (convergent validity) of the individual methods using G-theory as well as the more commonly used regression approach. Conventional regression analysis, based on the mean values for health states, revealed a coefficient of determination of 0.99, which is higher than the results achieved by Torrance (1976).

Our experiment confirmed another phenomenon found in earlier studies, i.e., SG_{ind} yields slightly higher valuations than does TTO_{ind}. Risk aversion may be one of the explanations. Risk aversion is assumed to lead to a relatively higher valuation of intermediate outcomes (health states) (Wakker & Stiggelbout, 1995); in this study, the bad health states in particularly were valued higher by SG_{ind} in comparison with TTO_{ind}.

The individual generalizability coefficient for the TTO method of 0.64 slightly exceeded the 0.56 obtained for SG. The inherent complexity of the SG task and the fact that people generally have difficulty with probabilistic situations may have been responsible for this result. The main method effect (SG versus TTO) was nil. This result of G-theory analysis was equal to the results of the regression analyses, which showed predictions up to 99%. Another conclusion is that the influence of interactions between health states and methods proved to be almost zero. This means that under the conditions of this study there were no systematic differences between the ordering of the valued health states with the two tradeoff methods. Moreover, it is clear that certain participants deviated from the group when valuing particular states, regardless of the method involved. The results do not support the claim for the unique conceptual position of SG, although a specific small effect, probably due to risk, could be observed.

We conclude that under highly controlled experimental circumstances the SG and the TTO are equivalent to a large extent, despite their apparent conceptual difference. Our results can be compared with the few existing studies that have examined this issue, taking into consideration that in the latter studies the numbers of health states and/or participants have usually been small and the statistical techniques rather global. The authoritative paper of Torrance published in 1976, reported a reliability coefficient (Pearson correlation based on replications) of 0.77 for both SG and TTO and a coefficient of determination (R^2) of 0.95 between SG and TTO. These coefficients are based on the mean values of six health states. In Torrance's study, the very bad and the very good health states were excluded, which may have improved the coefficients. Comparison of mean values obtained with the SG and the TTO for 35 disability levels by Wolfson *et al.* (1982) resulted in an R^2 of 0.84. Read *et al.* (1984) presented a Pearson correlation coefficient of 0.65 between the SG and the TTO. Their study was based on the valuation of only two health states. Hornberger *et al.* (1992) reported a Spearman rank correlation of 0.31 between the SG and the TTO. Their results were based on 58 individual patients' valuations of their own health.

Two issues arising from this study need some clarification. The first concerns the relatively high correlation coefficients that we found at the group level despite the considerable variation among the individual values. This can partially be explained by the fact that valuing health states is a stimulus-scaling task (Froberg & Kane, 1989b). There is to some extent a logical dominance of health states, which diminishes the variation among participants. For example,

all participants value 11122 as better than 33321. In contrast to attitude questionnaires, stimulus-scaling tasks frequently show high reliability coefficients. The wide range of the health states chosen as stimuli is another factor responsible for the rather high correlation coefficients.

The second issue refers to the measurement level (Stevens, 1946). The purpose of most health-valuation studies is to provide a valid representation of health states along a single continuum (construct) with specified anchors for a specified population. The individual values for our type of data are neither interval data nor ordinal data but more likely to be something in between, i.e., "quasi-interval" data. Individual values of health states are at best seen as imprecise representations of the individuals" internal scales of health states. An index of this imprecision, the generalizability coefficient (internal consistency) of individual values can be seen in Table 5. Information about the imprecision of individual measurements and the considerable differences among individuals is lost if we use the mean (Table 5) (Torrance, 1986). For applications where we are interested in group values, this seems an appropriate approach. Aggregation of individual quasi-interval outcomes results in a group scale that has real interval characteristics (Kerlinger, 1986).

We conclude that the feasibility of the SG and the TTO were comparable in this setting, while the reliabilities or precision of both methods based on the responses of all 104 participants were excellent. Is there any basis for preferring either method? Examination of the issue of content validity is difficult, as there is no agreed-upon "gold standard". The results of our studies have replicated the equivalence of two of the best-known tradeoff methods for the valuation of health states, the SG and the TTO.

Further research should particularly address such issues as the characteristics of the two tradeoff procedures in relation to risk attitude (Nease, 1994), time preference (Gafni, 1995), and other effects that influence the outcomes of these two elicitation methods as well as others (Loomes & McKenzie, 1989). Within the field of psychometrics, a major issue is how both methods allow for the use of other non-extreme reference states in the valuation task, elaborating on the axiomatic work of Llewellyn-Thomas *et al.* (1982).

We found considerable evidence that collective responses are comparable to the medians of individual responses for these two tradeoff techniques, although worse health states were valued lower under the collective response mode. We therefore consider it to be valid, especially for the moderate health states, to use the median of individual valuations as measure of the valuation of the group.

Acknowledgment

The authors thank the members of the EuroQol Group, in particular Paul Kind and Stefan Björk, for their constructive comments on the design of the study and on an earlier version of this paper, and Rosalind Rabin for her refinements to the manuscript. Valuable comments were also made by two anonymous reviewers and Anne Stiggelbout, PhD.

Appendix 3.1

Time-tradeoff (TTO)

In the standard TTO method, subjects are asked to judge a duration Y of survival such that surviving Y years in perfect health (Q^*) followed by death would be equal in preference to surviving, for example, ten years in a designated health state Q followed by death. The equation for utility in a standard TTO is:

$$U(X, Q) = U(Y, Q^*) \tag{1}$$

where $U(Q^*)$ ("best imaginable health state") is set at equal to 1 and X is 10 years,

$$U(Q) = Y/10. \tag{2}$$

In the present study we used "worst imaginable health state" instead, fixing $U(dead)$ at 0 as in the standard TTO. The utility of this reference state $U(W)$ can be set at 0 (depending on the theoretical assumption of "dead"), or it may be assumed to be known, as it is not measured directly. If assumed unknown, equation 2 becomes, for our study:

$$U(Q)=\left(\frac{Y}{10}\right)+\left(10-\frac{Y}{10}\right)\times U(W). \tag{3}$$

Equation 3 shows that the ratios $Y/10$ of standard TTO differ from the utilities of health states by an unknown additive factor $U(W) \times (10 - Y)/10$. Utilities elicited in this study are therefore not comparable with utilities elicited by standard TTO in other studies, unless $U(W) = 0$.

Standard-gamble (SG)

In the conventional SG, subjects are asked to choose a probability P such that a P change of surviving a lifetime in perfect health (Q^*) and a $1-P$ chance of immediate death would be equal in preference to surviving a lifetime in health state Q:

$$U(Q) = U(P, Q^*) \tag{4}$$

where Q^* ("best imaginable health state") is equal to 1,

$$U(Q) = P. \tag{5}$$

In this study "worst imaginable health state" replaced "dead" as the worst health state. An undetermined additive constant has thus been introduced, and equation 5 becomes:

$$U(Q) = P + (1-P) \times U(W)\,. \qquad (6)$$

If the (individual) assessment of the health state "worst imaginable health state" for the non-standard SG and the non-standard TTO are equal, equations 3 and 6 will both contain an unknown additive factor that is the same for the two methods. Comparisons and tests of equivalence with the SG and the TTO and their collective counterparts are therefore still permitted and justified.

Notice that constant $U(W)$ is not introducing an additional error component into the outcomes of our study, compared with the conventional SG and TTO. Setting $U(dead)$ equal to zero is an arbitrary convention, not precluding variation of individual assessments for the health state "dead" due to unique individual attitudes towards death itself.

Appendix 3.2

Classic Test Theory (CTT)

The true-score model is the core of CTT and is expressed as:

$$X = T + E \qquad (1)$$

X is an observed score that could be envisaged as the composite of two hypothetical components: a true score (T) and an undifferentiated random-error component (E). The correlation coefficient that expresses the degree of relationship between true and observed scores is known as the reliability index:

$$\rho_{XT} = \frac{\sigma_T}{\sigma_X} \qquad (2)$$

This coefficient can be easily estimated in CTT if the data of k related items are available by computation of a coefficient of internal consistency (also referred as homogeneity, scalability, or the like); Cronbach's α:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_\rho^2}{\sigma_X^2}\right) \qquad (3)$$

With multi-item health-status questionnaires, the goal is usually to measure a specific position

of a person on a domain. The question to be answered by Cronbach's α is how well the composite of all the items (σ_x^2) of a test is measuring the construct/domain.

Assessment of health states with an underlying natural ordering (stimulus-scaling task) yields another type of data. Instead of persons, stimuli (health states) are positioned on a scale. The question now is: how well does the composite of the responses of all the persons to the stimuli represent the scale of these stimuli?

An adaptation of the conventional Cronbach's α is consequently required. It is not the items (health states) that are tested for their there internal consistency, but the responses of the persons to the stimuli. Therefore, equation 3 shows index p (persons) instead of i (item).

Generalizability (G) Theory

A much more flexible framework in comparison with CTT is G-theory, fully based on analysis of variance (ANOVA). In this study we have used G-theory to estimate the generalizability coefficients (internal consistencies) for SG and TTO, which resembles, in this case, equation 3 while there are two effects (health states, persons). The object of the measurement is the valuation of the health states. Moreover, we were able to estimate generalizability coefficients based on scores of the responses of all 104 persons. The computational formulas and expected mean squares for estimating the sources of variance and generalizability coefficients for Table 4 are shown in Table B1.

Table B1

Source of variation	Mean square	Expected mean square	Estimated variance component
Health States (h)	MS_h	$\sigma^2_{hp,e} + n_p\sigma^2_h$	$\hat{\sigma}^2_h = (MS_h - MS_\varepsilon)/n_p$
Persons (p)	MS_p	$\sigma^2_{hp,e} + n_h\sigma^2_p$	$\hat{\sigma}^2_p = (MS_p - MS_\varepsilon)/n_h$
$h \times p, e(\varepsilon)$	MS_ε	$\sigma^2_{hp,e}$	$\hat{\sigma}^2_{hp,e} = MS_\varepsilon$

$$E_{\rho^2(individual)} = \frac{\sigma^2_h}{E\sigma^2(X)} = \frac{\sigma^2_h}{\sigma^2_h + \sigma^2_p + \sigma^2_\varepsilon};$$

$$E_{\rho^2(group)} = \frac{\sigma^2_h}{\sigma^2_h + (\sigma^2_p / n_p) + (\sigma^2_\varepsilon / n_p)}.$$

ANOVA estimates of variance components are unstable, especially with small sample sizes, and may even be negative. ANOVA is also sensitive to distribution form and unbalanced design. Our data failed only insofar as not all valuations of the health states showed normal distributions. This failure to meet the assumptions of ANOVA tends to overestimate error variances. Therefore, outcomes of the G study are conservative and may underestimate the true reliability coefficients.

Chapter 4

The Comparability and Reliability of Five Health-State Valuation Methods

Paul F.M. Krabbe, Marie-Louise Essink-Bot, Gouke J. Bonsel
Social Science and Medicine; 45: 1641-1652, 1997

Abstract

The objective of the study was to consider five methods for valuing health states with respect to their comparability (convergent validity, value functions) and reliability.

Valuation tasks were performed by 104 student volunteers using five frequently used valuation methods: standard-gamble (SG), time-tradeoff (TTO), rating scale (RS), willingness-to-pay (WTP) and the paired comparisons method (PC). Throughout the study, the EuroQol classification system was used to construct 13 health-state descriptions. Validity was investigated using the multitrait-multimethod (MTMM) methodology. The extent to which results of one method could be predicted by another was examined by transformations. Reliability of the methods was studied parametrically with generalizability theory (an ANOVA extension), as well as non-parametrically.

Mean values for SG were slightly higher than TTO values. The RS could be distinguished from the other methods. After a simple power transformation, the RS values were found to be close to SG and TTO. Mean values of WTP were linearly related to SG and TTO, except at the extremes of the scale. However, the reliability of WTP was low and the number of inconsistencies substantial. Valuations made by the RS proved to be the most reliable. Paired comparisons did not provide stable results.

In conclusion, the results of the parametric transformation function between RS and SG/TTO provide evidence to justify the current use of RS (with transformations) not only for reasons of feasibility and reliability but also for reasons of comparability. A definite judgment on PC requires data of a complete design. Due to the specific structure of the correlation matrix which is inherent in valuing health states, we believe that full MTMM is not applicable for the standard analysis of health-state valuations.

Key words: valuation methods, validity, comparability, reliability, methodology, EuroQol

Introduction

It is no longer sufficient to gather data on mortality and medical disease-specific parameters to evaluate the burden of disease and the effects of medical interventions from a societal perspective. Data on economic costs and on health status is also required. Considerable effort has been invested in the development of general indicators which allow for valid comparison of health-status effects across different diseases. It is generally agreed that such indicators should be derived from a comprehensive concept of health, covering at least the physical, psychological and social domains. Several indicators are currently available, some of them suitable for use in economic cost-utility analysis.

The following three-stage procedure is frequently used to incorporate health-status effects in utility analyses (Brooks, 1995; Essink-Bot, 1995). In stage I, the course of a disease is divided into broadly homogenous phases and patients' health status in each phase is measured using a descriptive system. In stage II, the health-status descriptions that correspond to the disease phases are formally valued. Results from stage I and II can then be combined with duration data in stage III to calculate quality-adjusted life years as an outcome measure.

The valuation of health states (stage II), forms a critical part of this three-stage approach. Several valuation methods (methodologically labeled: scaling methods; Froberg & Kane, 1989a) exist, each with their own theoretical framework and conceptual position. We investigated five established health-state valuation methods. First, we looked at a common rating scale, a seemingly simple method. Second, we investigated two economic methods, standard-gamble (considered to be the approximate operationalization of game theory) and willingness-to-pay, each referred to as tradeoff methods. From an economic point of view willingness-to-pay can be considered to be the superior quantification of non-monetary aspects of disease (Thompson *et al.*, 1982; Thompson *et al.*, 1984; Gafni, 1991; O'Brien & Gafni, 1996). We also investigated another tradeoff method, the time-tradeoff. This method occupies a position in between, i.e., being considered as more feasible than standard-gamble and more "realistic" than the rating scale. As a fifth method we added paired comparisons, a common psychometric indirect scaling method. Paired comparisons is considered to be the best scaling method from a cognitive point of view. It is based on less complicated binary choices instead of the direct assessments that are required for the other four methods. Paired comparisons is based on measurement theory (Torgerson, 1958) and was used in one of the first studies which focused on the elicitation of valuations for health states (Fanshel & Bush, 1970).

Throughout the experiment, the EuroQol classification was used and all the design features of the EuroQol valuation questionnaire were applied, except those related to the valuation technique (EuroQol Group, 1990). The generic EuroQol descriptive system for health states is suitable for all valuation methods and has been used extensively in fundamental and applied valuation research (EuroQol Group, 1990; Essink-Bot *et al.*, 1993; van Agt *et al.*, 1994; O'Hanlon *et al.*, 1994; Selai & Rosser, 1995).

A few studies have focused on to the simultaneous comparison of more than two methods (Torrance, 1976; Bombardier *et al.*, 1982; Llewellyn-Thomas *et al.*, 1982; Sutherland *et al.*, 1983; Read *et al.*, 1984; Hornberger *et al.*, 1992; Bass *et al.*, 1994; O'Brien & Viramontes,

1994). Most studies only partially standardized the stimuli and the testing conditions, hampering the interpretation of interstudy differences and preventing replication. In the experimental study described here we have tried to pay close attention to differences caused by the methods themselves instead of unintentional local conditions.

Most of the theoretical assumptions underlying the current valuation methods, though tenable, have yet to be empirically proved and there is evidence that some of the assumptions need adjustments (Johannesson *et al.*, 1994; Verhoef *et al.*, 1994; Gafni, 1995; Wakker & Stiggelbout, 1995; Bleichrodt, 1996; Stalmeier *et al.*, 1996). However, the present study is not oriented towards the testing of the underlying assumptions of the five methods. This paper essentially focuses on two questions: 1) to what extent do the five valuation methods yield comparable results, and 2) which of the methods is statistically the most reliable?

The first question deals partially with the validity of the methods. Validity encompasses three main aspects each with a rather broad scope: content validity, criterion-related validity and construct validity. Content validity refers to the question: "Is the instrument really measuring what we intend to measure?" For the purpose of this study, this implies a discussion about the "real" meaning and interpretation of values elicited by valuation methods. Are they really representing individual expressions of health-state preferences? Criterion-related validity is only applicable if one method can be identified as superior, i.e., a "gold standard". As these issues are part of an ongoing debate (Froberg & Kane, 1989b; Nord, 1992), content and criterion-related validity were not investigated directly in this study. Here we are primarily dealing with convergent validity which may be regarded as a type of construct validity. Convergent validity was studied by examining equivalence and comparability. First, we investigated the equivalence of the valuation methods, e.g., are particular health states *absolutely* valued equally by different valuation methods? Second, we investigated comparability, a broader concept related to the *relative* relationship between valuation methods. Equivalence was tested by comparisons of raw values, comparability allows for (restricted) transformation of data (e.g., value functions).

As part of a recently proposed standard approach to the comparison of methods (Streiner & Norman, 1995), we studied the different sources of measurement error which enabled us to reveal the reliability of the valuation methods in detail.

Material and Methods

The health-state descriptions

For the description of health states we used the classification developed by the EuroQol Group (Brooks, 1996). The EuroQol classification describes health status according to five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each dimension has three levels, i.e., "no problems" ("1"), "some problems" ("2"), "severe problems" ("3"). "Holistic", "comprehensive" (Froberg & Kane, 1989a) health-state descriptions are composed by taking one level for each dimension (e.g., the best health state is represented by 11111). Theoretically this set of dimensions and levels of the EuroQol

instrument allows for 243 (3^5) different health-state descriptions ("vignettes"). The EuroQol Group selected 11 of these vignettes as a standard set for experiment and study. Two health states were added in the present experiment as previous results indicated that the original set did not evenly cover the continuum between 0 (zero) and 100. Within each separate experiment, the vignettes to be valued were presented in a randomized order to avoid memory effects.

Short description of the five valuation methods

Rating Scale (RS) — The RS used in this study was the EuroQol "thermometer". This rating scale is presented as a vertical thermometer with a scale from 0 to 100. The anchors were labeled "best imaginable health state" at the top of the thermometer (100) and the "worst imaginable health state" at the bottom (0). The participants' task was to locate 13 different vignettes on a scale between the two anchors to represent their preference, in such a way that the intervals between the positions of the vignettes corresponded to the differences in preference as perceived by the participant. The task required the respondents to locate all the vignettes on the same scale.* This scaling task differed from the other three direct valuation methods (standard-gamble, time-tradeoff, willingness-to-pay), which required the health states to be valued separately.

Standard-Gamble (SG) — The SG method comprises an iterative paired comparison. SG involves making choices under conditions of uncertainty. Participants have to reach a point of indifference between the two alternatives by varying one of them using a "ping-pong" strategy. Conventionally, SG is operationalized as a choice between being in a specific lifelong stationary impaired health state (the state to be valued) or a hypothetical procedure with two outcomes: a probability (p) of instantaneous and lasting improvement to perfect health; or a complementary probability ($1-p$) of, conventionally, immediate death. By varying the p-level of the uncertainty outcome, the point of indifference between the two alternatives is determined. The value (utility) of the stationary state is defined as the probability p at the indifference situation. SG_c is regarded as a valid operationalization of the von Neuman-Morgenstern utility gamble (von Neumann & Morgenstern, 1953; Cohen, 1996).

The descriptions of the best outcome, generally described as "perfect health" or "optimal health" in other studies, was phrased as "best imaginable health state" in the present study. Our operationalization of SG differed from most publications in the choice of the lower anchor point in the gamble. We replaced "being dead" with the "worst imaginable health state" primarily for reasons of standardization between methods. This choice can be justified based on the assumptions of the method (Llewellyn-Thomas *et al.*, 1982; Torrance, 1986). It was clearly stressed to the participants that both outcomes arising from the gamble would involve chronic health states. Values obtained in this way require a linear rescaling factor to be comparable with values obtained with the standard SG, assuming perfect scalability of "dead"

* Usually the application of RS implies that for each stimulus valuation a separate rating scale is used. Here the health states were valued simultaneously in two sets of vignettes on facing pages, with on each page a vertical scale.

and the "worst imaginable health state" on the assumed health continuum (Krabbe *et al.*, 1996).

Time-Tradeoff (TTO) — This method was developed by Torrance as a less complicated, conceptually different although equally sound alternative to SG. Like SG, TTO is based on trade-offs, but the concept of uncertainty is omitted. Participants trade off survival time and health status. In the conventional operationalization, the first alternative offers a (sub optimal) stationary health state with a given duration (x), 10 years in the present study. A better health status (conventionally perfect health) of shorter duration is offered as the competing alternative, conventionally followed by death. The point of indifference is reached by varying the duration spent in perfect health (y). Subsequently, by combining x and y, the value of the stationary health state can be established (y/x). For reasons already mentioned, we replaced "being dead" by "worst imaginable health state" in the present study (Krabbe *et al.*, 1996). The optimal health state was phrased as "best imaginable health state". For both options the health state would return to its present form after ten years.

Willingness-To-Pay (WTP) — The willingness-to-pay task in our study started by confirming the average budget situation of the medical students participating in the experiment. A monthly budget of $ 725 (standard study grant of 1200 Dutch Florins given by the Government, 1993) could be spent after subtraction of the rent for a room and fixed costs for food, heating, clothing etc. (500 Dutch Florins). Respondents were asked to imagine that they were in a certain impaired state of health and asked what amount they were willing to give up permanently to return to their previous (healthy) condition. This operationalization was chosen after piloting available alternatives.

Paired Comparisons (PC) — PC is a scaling method consisting of a two-step procedure (McIver & Carmines, 1981). PC is especially developed for scaling unconcatenate subjective attributes (such as: food, politicians). As in the three trade-off methods, the participant is confronted with two outcomes, but here preference is required rather than trying to achieve a point of indifference. The data on individual preferences between all possible pairs of health-state descriptions allow for the construction of a matrix of ½(n(n–1)) preferences, expressed as probabilities. The probability in every cell of the matrix is the proportion of the "row" health state being preferred to the "column" health state by the judging panel. As a second step, transformations and computations based on scaling theory, construct a unidimensional interval scale of health states.

In this experiment, the 13 health states to be scaled, comprised a considerable number of paired comparisons consisting of so-called dominant pairs, i.e., one of the two health states is objectively "by definition" better than the other health state (e.g., "12232" is more severe than "12132"). Out of 78 possible pairs† 43 pairs were dominant. For reasons of efficiency only the remaining 35 non-dominant pairs (45% of all the possible pairs) were valued.

After the standard forced choice comparison, we requested a graded choice (scale 1 – 9: 1 = strong preference health state A, 5 = indifferent, 9 = strong preference health state B), (see also: Hadorn *et al.*, 1992).

---

† $\frac{n(n-1)}{2} = \frac{13\times(13-1)}{2} = 78.$

Organization and testing conditions

The experiment included two sessions, separated by a 10-day interval. The same group of 104 students participated in both sessions. Students were recruited by handouts. For full participation they were paid a fee of approximately $ 65 (1993). Data collection took place in a group, since another objective of the experiment was to study the equivalence of collectively and individually collected responses (published elsewhere, see: Krabbe *et al.*, 1996). Both sessions consisted of a sequence of valuation tasks deliberately interspaced with unrelated questionnaires, e.g., on the moral acceptability of genetic manipulation. From pilot studies with other participants we learned that weariness and even irritation due to monotony had to be prevented by alternation of tasks and the inclusion of breaks.

All participants were seated in a lecture hall with due space between them. Each different method was preceded by a similar verbal explanation of the method and a few test judgments. The descriptions of the health states to be valued were always presented by slide projection. During the presentation, the instructors (GJB, MLE-B) repeated the nature of the particular method for each valuation to avoid blurring of the concepts.

Values for the methods RS and WTP were elicited during the first session, SG and TTO during the second session. The collection of the PC method responses were divided over the two sessions (both PC forms alternate for each health state).

Responses were collected by pencil and paper for RS, SG, TTO and WTP, and by means of an electronic response system (standard PC: choose A or B) for the two types of PC methods. For SG each participant responded by dividing a "probability pie" into two complementary parts. The task for TTO was to divide a "duration bar" into two parts.

All the separate tasks were pretested with other panels and a detailed work schedule was used to ascertain equivalence of presentation, of explanation, etc.

In order to detect differences associated with characteristics of the methods themselves, we controlled for the following:

- factors related to the health states (such as prognosis) were kept constant;
- factors related to the subjects who performed the valuation tasks (age, education, experience with illness) were kept constant by selecting a homogenous panel;
- characteristics of presentation of the health-state descriptions (order, framing, lay-out, instructions) by written protocols and training.

Analysis

Outcomes of the RS, SG, TTO and WTP methods were transformed by linear transformation to a uniform 0 to 100 scale (RS = score; TTO = 10 × score; SG = 100 × score in degrees/360; WTP = [Dfl. 1200–Dfl. 500–score]/7).

To analyze the partial preference matrix of the PC task we used Thurstone scaling (Torgerson, 1958; Hadorn *et al.*, 1992) to derive a unidimensional, interval scale of health state preferences. For the graded paired comparisons task, we computed the average preference rating (APR) as described by Hadorn *et al.* (1992). We included all the responses for each of the five methods, although they were not fully complete due to missing values (see below).

If a valuation method is cognitively easy to handle and clear to understand (feasibility), it might be expected that in dominant pairs of health states, the better state is preferred. If this is not the case the results are viewed as inconsistent. In order to study inconsistencies in the valuation of dominant pairs, we computed distances between the dominant health state and the secondary health state for all relevant pairs. According to our definition, the distance between vignettes "33332" and "11112" is the summation of the level differences for the five dimensions. For this example the distance is: (3–1 =) 2 for dimension one-four and is (2–2 =) 0 for the last dimension, yielding a total distance of 8. Vignette "33332" had the largest distance in relation to vignettes "11112", "11121", "11211", "12111", and "21111". Respectively the smallest distance was between vignette "11122" and two vignettes "11121" and "11112".

Validity — simple. Convergent validity between the methods was investigated by Pearson's product-moment correlation coefficient and Spearman's rank-correlation coefficient. The first statistic is suitable for interval or ratio data while the second statistic is more appropriate for ordinal data or for data of higher measurement level that does not satisfy requirements for ρ. To test exact concordance of continuous data, we also computed intraclass correlation coefficients (ICCs). ICCs include level-effects between different measurements. These three coefficients for convergent validity were all computed based on the mean values for the 13 health states.

Validity — construct. To study construct validity for the four direct valuation methods (PC could not be included being an indirect scaling method, yielding a different type of data) we applied the multitrait-multimethod methodology (MTMM) on the individual responses (Crocker & Algina, 1986; Hadorn & Hays, 1991). Based on a matrix representing all the intercorrelations between multiple traits (13 health states) and multiple methods (RS, SG, TTO, WTP), four classes of correlations can be distinguished (see: Appendix 4.1).

Validity — convertability. We examined the numerical comparability among the methods. If valuation methods are not equivalent (i.e., they do not give the same values for instance, intraclass correlations coefficients are not high), perhaps values are related in some systematic way so that conversion curves can be constructed. Power functions (method Y = 1–[1–method X]$^{\alpha}$), similar to Torrance (1976), were therefore estimated relating mean values of the 13 states for all six pairs of methods. Computations have been performed by the non-linear regression module of SPSS for Windows.

Measurement error/Reliability. We used Generalizability Theory (G-theory) as a general approach to estimate the relative contribution of the multiple sources (facets in G-theory language) to measurement error/bias (Streiner & Norman, 1995; Krabbe *et al.*, 1996). G-theory is a specific application of analysis of variance (ANOVA) and requires individual data. In the present case, the relative contribution (variance components) of the facets "health state", "method" and "participants", their interaction terms as well as all other facets of measurement error, were estimated separately.

Furthermore, G-theory was used to estimate reliability coefficients for the separate methods. These reliabilities are closely related to the internal consistency concept (Cronbach's alpha). Although instead of the stimuli (health states), here the responses of the participants to the stimuli as they were elicited by the valuation methods were tested.[‡] Hence, agreement among participants was estimated rather than similarity of items (health states). G-theory is a

method which treats valuations at interval measurement level. In order to study the internal consistency reliability among the participants in their valuation of the set of 13 health states, but treating the valuations as rankings, Kendall coefficients of concordance W were determined (Siegel & Castellan, 1988), concurrent with the G-study.

Results

Response

Of the 104 participants in this study, 46% were male. All were students, 71% were medical students. Mean age of the group was 22 (SD = 2.48) years. RS, SG, TTO and WTP each took about 15 minutes to complete. The responses of all 70 PC pairs (35 standard and 35 graded) using the voting system took about 90 minutes. Judging from participants' remarks and from the absence of learning effects, we regarded memory effects to be highly unlikely. The feasibility of these experiments was satisfactory, although at the end some participants complained of weariness. Few responses were missing.

Descriptive statistics and consistency measurement

The results of the paired comparisons method proved to be unstable. If the empirical data were changed in only a minor way this resulted in a major alteration of the unidimensional scale. The background to this was the decision to leave out the apparently dominant pairs from the empirical task, leaving the relatively difficult ones to be measured and scaled. Empty cells were substituted with "expected" preferences, but this approach also yielded unstable results and was therefore rejected. Thus no results from the PC method will be presented in this paper.

Table 1 shows the results (means and standard deviations for each health state and overall means and ranges for the methods) of the experiment for the four remaining methods. The order of presentation of the 13 health states in Table 1 is arbitrarily based on the SG values. Mean valuations for the 13 health states for the four methods are shown in Figure 1.

A summary of the measures for inconsistency (individual level) is presented in Table 2. As expected, the inconsistency was highest for the pairs with the smallest distances. Average inconsistency for the methods SG and TTO was almost the same (4.6% respectively 4.3%). In RS it was lower (2.0%) and in WTP higher: 7.4%. We observed a 50.5% inconsistency for the method WTP for the two dominant pairs with distance 2.

‡ The reason for this exchange is that we are dealing with data that stems from a so-called stimulus-scaling task (see: Froberg and Kane 1989b).

Table 1
Mean values and standard deviations for the 13 health states (n = 97..104; between parentheses SG order) by the four methods (all linearly transformed* to 1 – 100)

Health State**	Standard-Gamble		Time-Tradeoff		Rating Scale		Willingness-To-Pay	
	mean	SD	mean	SD	mean	SD	mean	SD
12111 (1)	96.2	5.3	94.4	8.8	68.2	12.8	89.5	9.1
11211 (2)	95.6	4.4	92.6	7.7	73.4	11.7	86.8	11.2
21111 (3)	94.5	7.0	92.8	8.7	71.7	10.1	83.1	14.6
11112 (4)	93.8	12.3	93.6	8.0	70.3	11.8	87.2	13.9
11121 (5)	93.3	8.9	91.8	8.6	70.2	12.7	86.4	14.6
11122 (6)	89.0	13.5	86.0	11.0	55.0	12.0	81.1	15.6
12212 (7)	81.7	15.1	78.6	14.4	47.0	12.7	75.9	13.9
32211 (8)	79.2	18.6	73.1	18.6	41.2	12.8	65.3	19.1
21232 (9)	65.2	22.8	59.0	20.3	31.1	14.2	60.1	19.2
22323 (10)	64.5	23.7	61.0	22.6	24.6	13.0	59.5	17.5
33321 (11)	53.6	26.5	47.8	24.0	26.4	12.7	58.6	18.4
22233 (12)	51.5	28.4	44.9	24.7	22.1	13.4	52.0	18.4
33332 (13)	34.4	25.3	27.8	23.4	10.7	9.0	45.6	18.9
Range Means	61.8		66.6		62.7		43.9	
Mean SD		16.3		15.5		12.2		15.7

* SG, scores transformed as: SG = 100 × score in degrees/360; TTO, scores transformed as: TTO = 10 × score; RS: untransformed scores; WTP, scores transformed as: WTP = (700 - $WTP_{original}$)7.

** The EuroQol concept of health status consists of five dimensions: mobility, self-care, usual activities, pain/discomfort, and mood. Each dimension has three levels, "no problem" (1), "some problems" (2), and "severe problems" (3). Health-state scenarios are produced by the selection of one level for each dimensions (the best health state is thus represented by 11111).

Validity

Figure 2 shows the correlations between the four methods as a first estimate of convergent validity. The Pearson product-moment correlation coefficients were high and close to 1.0 for all the six relationships between the four methods. Spearman rank correlations were slightly lower than the interval-based Pearson product-moment correlation coefficients. In Figure 1 the four lines do not match but are parallel. ICCs were much lower than the Pearson and Spearman correlations, particularly for the relationship between RS and the other methods, suggesting important level effects.

Due to specific patterns (see: Conclusions & Discussion) between the correlation coefficients of the data computed for the MTMM methodology, only a partial analysis was valid. This is presented in Table 3 which shows that convergent validity (monotrait-heteromethod coefficients)§ was reasonably good for all the health states across

§ Because these correlation coefficients are based on the analysis of individual values for the 13 health states of the four methods, such correlations are therefore always lower than (Pearson PM) correlation coefficients based on aggregated mean data (e.g., Figure 2).

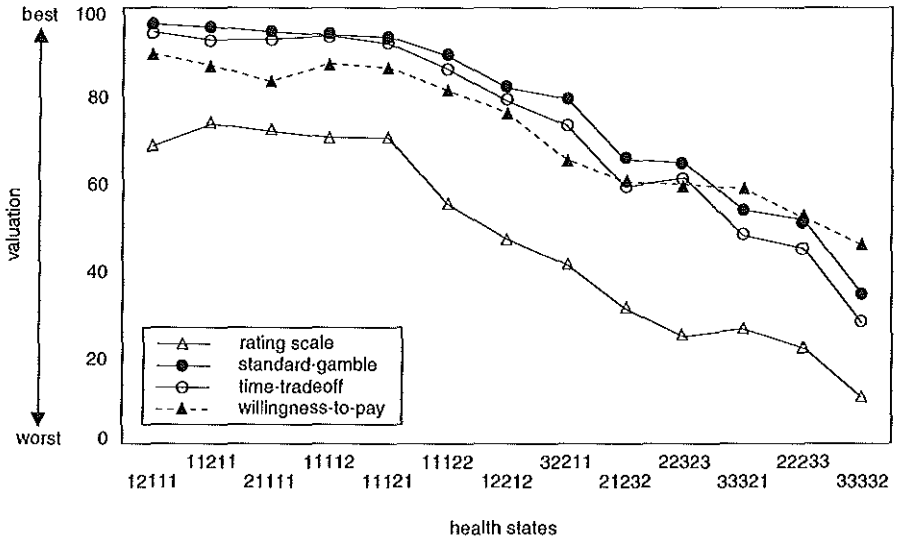


Figure 1
Valuations (means) for the 13 EuroQol health-state descriptions elicited by the four methods (ordered by the standard-gamble values)

Table 2
Percentages of inconsistencies between dominant pairs* of health states for each of the four method (n = 97 – 104)

Distance	Number of Pairs	Rating Scale	Standard-Gamble	Time-Tradeoff	Willingness-To-Pay
1	2	12.5	21.6	17.8	50.5
2	5	10.6	13.7	12.1	21.9
3	5	6.2	8.5	6.3	11.5
4	7	3.2	7.1	5.1	7.8
5	3	2.6	4.8	5.4	4.5
6	15	0.4	2.9	3.0	6.3
7	1	1.0	0.0	1.9	1.9
8	5	0.0	1.9	2.3	2.9
Total	43	3.4	6.3	5.5	10.6
Weighted Total**		2.0	4.6	5.3	7.4

* Total number of valuable dominant pairs: 43 × 104 = 4472.
** Weighted total: SUM (number of pairs × proportion of inconsistencies × N) / SUM (distance × weighted number of pairs).

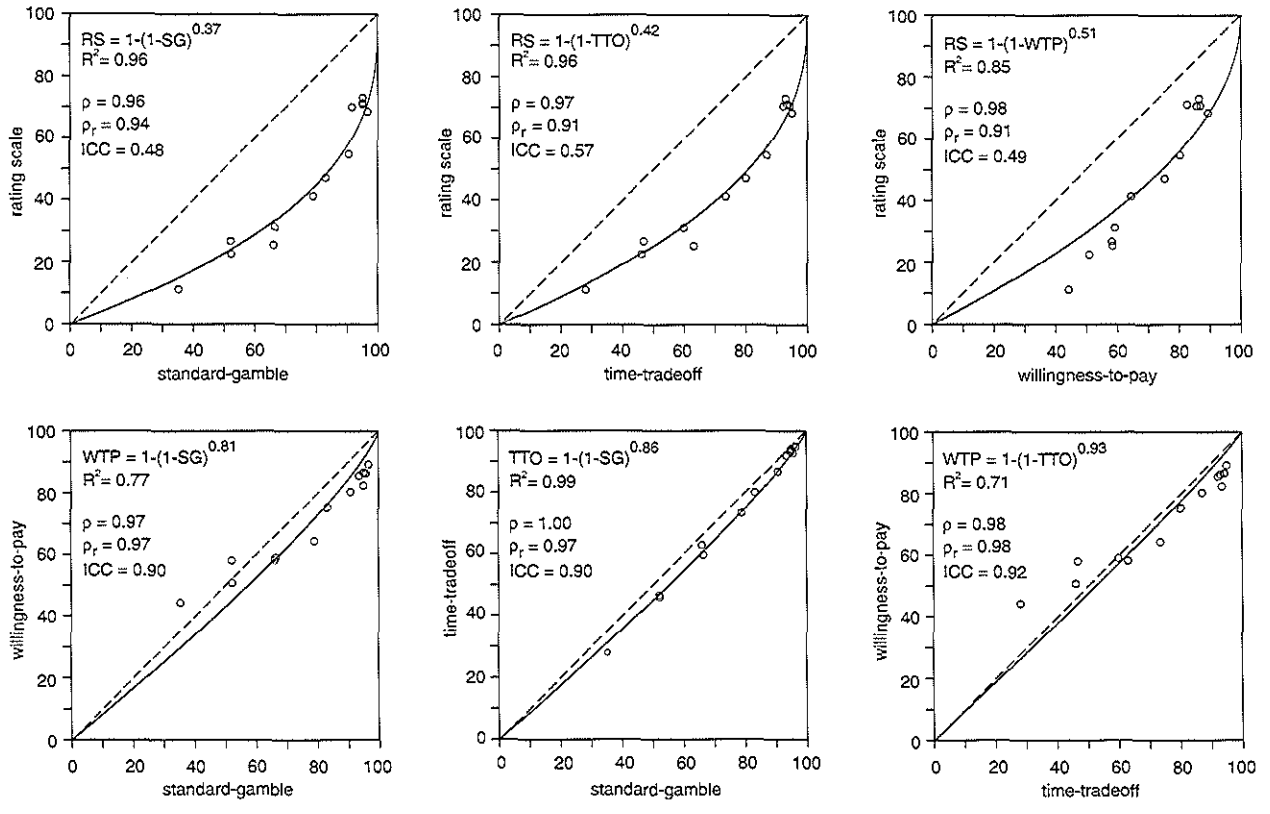


Figure 2

Convergent validity measured by correlation measures (ρ = product moment Pearson correlation, ρ_r = Spearman rank correlation, ICC = intraclass correlation) between the health-state values for the four methods based on the mean values (means of Table 1, n = 13) and functional relationships between the four elicitation methods studied by nonlinear regression analysis (power function); the entry labeled R^2 is the coefficient of determination and may be interpreted as the proportion of the total variation of the dependent variable around its mean that is explained by the fitted model

Table 3

Convergent validity (monotrait-heteromethod validity correlations = same health state assessed by different method) for the 13 health states and the four elicitation methods based on the individual values (n = 91)

Comparison	12111	11211	21111	11112	11121	11122	12212	32211	21232	22323	33321	22233	33332	Mean Methods**
RS vs. SG	0.46	-0.03	0.08	0.27	0.20	0.28	0.07	0.14	0.12	0.25	0.20	0.22	0.17	0.20
RS vs. TTO	0.36	0.07	0.22	0.19	0.20	0.33	0.20	0.19	0.18	0.33	0.20	0.20	0.10	0.23
RS vs. WTP	0.10	0.32	0.26	0.29	0.23	0.40	0.19	0.24	0.33	0.07	0.21	0.13	0.19	0.25
SG vs. TTO	0.63	0.66	0.72	0.59	0.84	0.77	0.52	0.65	0.67	0.74	0.73	0.70	0.67	0.69
WTP vs. SG	0.16	0.14	0.14	-0.02	0.29	0.16	0.12	0.13	0.07	-0.03	0.17	0.15	0.10	0.15
WTP vs. TTO	0.25	0.18	0.19	0.19	0.24	0.25	0.09	0.21	0.17	-0.04	0.12	0.11	0.07	0.18
Mean correlations between the four methods per health state	0.37	0.32	0.34	0.31	0.40	0.41	0.25	0.32	0.33	0.35	0.34	0.32	0.30	

* RS = rating scale, SG = standard-gamble, TTO = time-tradeoff, WTP = willingness-to-pay.

** For each health state and for each comparison between two methods the square root of the means of the squared correlation was computed to summarize the row/columns with correlation coefficients.

Table 4
Estimated variance components (percentages) of health states (13) × participants (91) × methods (4)

Source of Variation	Variance Components	
	Raw Scores	Individual scores transformed for each method to z-values
Health States (H)	44.8	72.0
Methods (M)	20.0	- *
Participants (P)	4.7	0.7
HM	2.3	2.0
HP	8.0	8.8
PM	7.1	1.2
Residual (HPM, ε)	13.1	15.3

* By definition.

the methods SG and TTO (overall: 0.69). All other comparisons between the four methods showed low convergent validity (overall: 0.15 – 0.25). Coefficients for the comparisons between tradeoff pairs WTP/SG and WTP/TTO were even lower than for RS/SG and RS/TTO. No particular pattern could be detected by inspection of the means of the convergent validity coefficients of the 13 health states in Table 3.

Except for WTP, values of all the other methods appeared to be exchangeable after a one-parametrical (α) power transformation. The percentage of variance explained by the power functions was: RS = f(SG), 96%; RS = f(TTO), 96%; TTO = f(SG), 99%. Accuracy of predictions including WTP was considerably lower: RS = f(WTP), 85%; WTP = f(SG), 77%; WTP = f(TTO), 71%. All plots of RS with the tradeoffs showed concave power functions (α = 0.37 through 0.51), the other functions were approximately linear (α = 0.81 through 0.93).

Measurement error and reliability

The initial analysis on the individual raw scores demonstrated that only 44.8% of the total variance was attributed to the variability of the 13 health states (H) scores (see: Table 4). A small percentage of variance, 4.7%, was accounted for by systematic differences in valuations of the health states by the facet participants (P). This relatively small contribution indicated that, averaged over all the health states and all the methods, the participants valued the health states only slightly systematically differently. Twenty percent of the total variance stemmed from the facet methods (M), which was in particular attributable to the divergent magnitude of the RS valuations. Overall, 17.4% of the total variance was attributable to the first-order interaction terms (HP, HM, PM). The interpretation of, for instance, the interaction term HP is that *some* participants valued *some* health states systematically differently. Only 13% of the

Table 5
Estimated variance components (percentages) of the health states (13) × participants (91) for each method separately

Source of Variation	Variance Components			
	Rating Scale	Standard-Gamble	Time-Tradeoff	Willingness-To-Pay
Health States (H)	77.0	57.6	64.6	48.9
Participants (P)	5.5	11.8	9.9	31.4
Residual (HP, ε)	17.6	30.6	25.5	19.7

total variance came from the three-way interaction HMP, which suggests a satisfactory explanation model for this data set.

Additionally the individual values for the health states were transformed to method-specific z-values. Absolute differences between the four methods (M), especially between the RS and the other methods, were eliminated by this standardization, yielding by definition zero variance for the factor methods (M) and increasing the percentage of variance uniquely attributable to the health states (H) to 72%.

The results of the G-study for each of the methods separately are shown in Table 5. For the WTP method more than 30% of the total variance was due to systematic differences between participants valuing the 13 health states, which was high compared to the other methods. Additional inspection of the data revealed that the relatively great contribution of systematic differences for WTP between the participants was reducible to two response patterns. One response pattern consisted of a small tradeoff of the budget except for the very severe health states, for which almost the whole budget was exchanged to remain in full health. The other response pattern showed exchange of almost the whole budget for even moderately bad health states (insensitivity for the stimulus, e.g., due to cognitive difficulty).

The variance components of the health states (H) can be regarded as (standard) reliability coefficients, assuming interval metric properties of the data. RS was the method with the highest reliability: 0.77. The reliability for WTP, 0.49, was low.

Nonparametric statistics revealed higher coefficients. For all the four methods there was good agreement among participants in their ranking of the health states (Kendall's W test; n = 91). The highest agreement, 0.83, was achieved by the RS method. Agreement using SG was sufficient: 0.75. For TTO and WTP the coefficients were 0.77 and 0.80 respectively.

Conclusions and Discussion

Under highly controlled conditions we conducted an experiment with five valuation methods. Design characteristics aimed at maximal standardization except for the two manipulated effects, i.e., health states and valuation methods. We assumed these to be responsible for the observed effects. Although we were able to control for many factors, other factors may still have influenced the results.

In this study, session effects are the most conceivable ones. Generalization of the study results may be further restricted due to the composition of the panel that performed the valuation tasks and to the selection of the 13 health states. A different selection of respondents could lead to different results, although several studies have shown that these effects in this context are minor (EuroQol Group, 1990; Essink-Bot *et al.*, 1993). The selection of the health states may have to some extent influenced the results of this study, although we expect the relationship between the methods to be hardly influenced because the sample of the health states was well chosen.‖

We will first clarify the unexpected outcomes of the WTP method and the problems that we confronted using the PC scaling method. We will then discuss the comparability of the methods and their reliabilities. Finally we will consider the complications we encountered when studying construct validity using the MTMM methodology.

We found it difficult to proceed with the WTP method, even among this homogeneous and highly educated population and despite our controlled study design with extensive explanation and test questions. Two typical response patterns appeared to determine the reliability and the range of the responses. Thus the WTP results were not satisfactorily comparable to the other tradeoff methods despite satisfactory regression results and inter-method comparability on first sight. Although a linear transformation of the mean WTP values to SG/TTO was technically possible, WTP in our operationalization was found to be an inferior method with an unacceptably low reliability. Even more worrying was the amount of inconsistency found between the dominant pairs of health states. Evidence from the few studies that have focused on WTP is difficult to interpret due to variability of concepts used, the small samples, and the small number of health states which do not allow for sound statistical testing (Thompson *et al.*, 1982; Thompson *et al.*, 1984; Thompson, 1986; O'Brien & Viramontes, 1994; Chestnut *et al.*, 1996). Unless it is possible to improve the operationalization of WTP it may have to be regarded as an unfeasible method. Perhaps therefore, the concept of WTP is only valid in real-life situations (sometimes called revealed preference or averting behavior method) and not suitable under experimental conditions.

‖ An alternative study based on a set of 13 health states selected by an restricted inclusion criterion (for example: 13 EuroQol health states, without level 3) would induce a decline of the proportion of variance (Table 4 and 5) for the facet "health states" and consequently yield lower reliability coefficients. However, we were not interested in the characteristics of valuation methods for a *specific* domain of health states. The selection of the health states was deliberately worked out to evenly cover the continuum between 0 and 100, which enables us to study the "behavior" of the participants for the valuation methods on the *whole range* of possible health states. In this context we are particularly interested in the comparison between the methods. The inclusion of a set of health states with a restricted range would have certainly decreased the proportion variance contributed to the health states, but would also obscure the division between the methods.

Serious problems were also encountered with the PC scaling method as it did not provide stable outcomes for both the PC variants (standard, graded). The underlying difficulty with the application of PC is the high number of pairs to be valued with a complete design and the probability of bias in dominant pair evaluation. Due to the partially ordered nature of our stimuli, we could not overcome the problem with the relatively high number of empty cells (dominant pairs) (MacCallum, 1978).

Hadorn *et al.* (1992) applied PC with a partial design with apparently more success (n = 93). In our analysis the factor critical to failure appeared to be the number of dominant pairs and the level of complexity of the classification. Hadorn *et al.* used only two dimensions (i.e., "pain or physical suffering" and "limitations on daily activities") with four levels each (EuroQol: five dimensions, three levels) and only selected relatively comparable pairs of health states. Therefore their PC analysis was based on an incomplete and selective design of 54 (59%) of the total number of pairs.[1] Reconstruction of Hadorn's design revealed that still 40% of these 54 pairs were dominant pairs (in our design 0%). Moreover, no mention was made by Hadorn *et al.* of the stability of the PC method for scaling of health states based on their incomplete design nor did they report the effect of the substitution of empty cells with "expected" preferences.

A surprising finding was the performance of the MTMM methodology in this context. The method was advocated by Froberg & Kane (1989c) for good reasons and empirically applied by Hadorn & Hays (1991). In retrospect, our failure with MTMM can be explained by the characteristics of data yielded by the process of valuing "subjective" stimuli such as health states as opposed to the more common situation where participants have to reveal their opinion on, for instance, the attractiveness of consumer goods with "latent" indivisible characteristics. In our study, health states have "manifest", ordered domains. If dominance exists, as is the case here, then the usual MTMM analyses are not adequate. Correlations between the health states then show a special structure indicated as a "simplex structure" (Jöreskog & Sörböm, 1979). The typical property of a simplex correlation matrix is that correlations decrease as one moves away from the main diagonal. Valuations of health states that were of the same severity will show moderate between-method correlation, but valuations of health states that were different in severity (e.g., "21111" vs. "33321") show no correlation at all (as was observed in our data). MTMM analysis requires at least moderate or low correlations among *all* health states elicited by one and the same method.

Hadorn and Hays (1991) presented an early application of MTMM analyses. Six aspects of health-related quality-of-life (HRQoL) were investigated (i.e., general health perception, physical suffering etc.). Participants (n = 76) were asked to provide preference ratings (valuations) by judging the effects of different levels of problems or impairments on each of the six dimensions on overall quality-of-life. This task was performed for three different assessment methods, developed by the authors themselves. As a result of their different

---

1 To the initial 16 constructed health states they excluded the two anchor states (no suffering-no limits, severe suffering-severely limited), which resulted into $\frac{14\times(14-1)}{2} = 91$ pairs of vignettes to be compared. The authors reported that despite the dominance restriction, 54 pairs remained to be assessed (after reconstruction of their design we arrived at 50 pairs). Therefore, at least 37 pairs were not valued as it was thought that one of the health states of such a pair was manifestly dominant.

strategy which was not dealing with the valuation of health states but with eliciting individual preferences for *separate* aspects of HRQoL, the problem of the simplex-structure that we encountered with MTMM was absent. After some consideration we judge MTMM incompatible with data analysis of standard n (independent domains) $\times$ p (ordered levels) classification systems.[¥]

We investigated the convertibility of the methods straightforwardly applying simple algebraic power functions. Torrance (1976) reported a power relationship $RS = 1-(1-TTO)^{\alpha}$ between RS and TTO with a coefficient of 0.62 ($R^2 = 0.80$) based on 18 means of valued health-state scenarios (n = approx. 200). In a study by Stiggelbout *et al.* (1996) a coefficient of 0.64 was presented. Loomes (1993) found a coefficient of 0.55 based on a secondary analysis of data by Bombardier *et al.* (1982). We found, based on 13 mean values, $\alpha = 0.42$ ($R^2 = 0.96$) for the power function. Busschbach (personal communication, 1996) reported similar results, namely $\alpha = 0.47$ ($R^2 = 0.95$; n = 103). Different coefficient values may be the result of many factors. Of the 18 scenarios in Torrance's study, none were valued very low or high, which may have caused the higher power coefficient. The study of Stiggelbout *et al.*, even more than Torrance's study, lacked a broad range of health states because each respondent valued only his/her own health state. Other factors that could be responsible for different outcomes are: the composition of the valuation panels, the instruction to the panel and the classification system used.

We conclude that valuations of health states based on rating scales are distinct from but strongly related to outcomes derived through tradeoff methods. Tradeoff methods elicit values expressing an individual's preference for a particular health state under a condition where something has to be *sacrificed* (e.g., change on good outcome, life years, budget). Rating scale methods however are based on the *comparison* of different health states. RS values express the subjects' internal representation of health states in a stable world where the actual health of the respondent probably plays a major role as a reference point.

The choice of which type of values is to be used depends largely on the perspective of application. From the individual perspective, generally directed at decisions on chance, tradeoffs seem more appropriate to elicit valuations. For collecting societally grounded health-state valuations the RS method presumably is a feasible tool, particular if ordering of health states is the restricted goal. Use in the context of societal decisions theoretically requires power transformation.

The reliability coefficients estimated by the G-study showed lower reliability for all the methods in comparison with Kendall's W concordance coefficient based on ranks. Reliability of a G-study takes not only the ordering of health states into account but also the distances between health-state values. This explains why, in the case of WTP, the G-theory reliability coefficient was only 0.49 versus the Kendall's W of 0.80. Reliability was satisfactory for SG

¥ The following two formulas relate to this topic. The number of pairs that can be achieved based on a descriptive system of p levels and n domains is $\frac{p^n(p^n-1)}{2}$. For computing the number of *dominant* pairs the formula is $\left[\frac{p(p+1)}{2}\right]^n - p^n$.

and TTO. In this study the RS method showed a reliability even higher than the two standard methods (see also: Torrance, 1986).

Taken together, a valid comparison of more than two valuation methods under highly controlled conditions is feasible and a simple power transformation suffices to describe the value function between health-state valuation methods. The RS method is in this sense almost congruent to SG and TTO.

Two interesting "negative" outcomes require further study. First, the PC method proved to be not applicable due to the dominant pairs of health states. Valuations of only non-dominant pairs of health states impairs accurate estimation of scale values. Inclusion of all pairs of health states yields highly flawed results. Also, the MTMM methodology appeared not to be suitable for essentially the same reason as the failure of the PC method.

Future consideration might be given to whether there are other techniques/ methodologies that are potentially valuable for the elicitation of valuations/preferences for health states. Unfolding analysis could be such a technique (Coombs, 1950; Lewis-Beck, 1995). It is fully focused on the analysis of preference data. Additionally, a methodology used with good results in a small number of fields is functional measurement (Anderson, 1976) and conjoint analysis (Louviere, 1988). A specific example of its implementation is the multiattribute application of Torrance *et al.* (1982). But most of all well structured experiments and studies are needed to clarify the numerous indistinct concepts and assumptions related to the use of health-state valuation methods.

Acknowledgment

The authors wish to thank the Advisory Board for the Health Research Promotion Program (Adviesgroep SGO) for financially supporting the Research Program "Standardization in Medical Technology Assessment". We would also like to thank C. Vonk, MSc. for providing the formula of the numbers of dominant pairs.

Appendix 4.1

Multitrait-multimethod methodology (MTMM)

Originally this methodology was introduced by Campbell and Fiske (1959). They identified four classes of correlation coefficients. First, monotrait-monomethod reliability correlations (health states measured twice for each method separately: test-retest). Second, heterotrait-monomethod correlations (different health states for the same method). Third, heterotrait-heteromethod correlations (different health states assessed by different methods). Finally, monotrait-heteromethod validity correlations (same health state assessed by different methods). Using MTMM, construct validity is supported if correlations among different methods are high for a single trait (convergent validity), but correlations between the same methods measuring different traits are low (discriminant validity). Although Campbell and Fiske recommended visual inspection of the MTMM matrix for assessment of construct validity, recent additional modeling procedures (e.g., confirmatory factor analysis) have been

developed which may lead to more unequivocal interpretation of such data (Schmitt & Stults, 1986; Jöreskog & Sörböm, 1989). We have performed analyses based on both classical Campbell and Fiske criteria and by using confirmatory factor analysis.

For the basic MTMM model based on confirmatory factor analysis, we treated each of the 13 health states as separate traits and the four valuation methods as separate methods. Another model was estimated by constructing three clusters of health states (mild, moderate, severe) as three separate traits. Models were also estimated assuming dependency (correlation) between the methods. For all models, various transformations (logit, arcsine, rescaling) of the data were used. However, none of these models led to meaningful outcomes as a consequence of the specific structure of the correlation matrix (see: Conclusions and Discussion).

Chapter 5

Sequence Effects, Health Profiles and the QALY Model:
In Search of Realistic Modeling

Paul F.M. Krabbe, Gouke J. Bonsel
Medical Decision Making; 18: 178-186, 1998

Reprinted with permission from Medical Decision Making

Abstract

The authors conducted an experiment to determine whether the sequence of presentation of states in a health profile would affect the valuations assigned to them. The empirical task was part of a large standardized experiment involving 104 students. Thirteen health states were valued using two variations of the time-tradeoff method.

At the group level, a small but distinct overall effect of the sequence of the tradeoffs was detected after accounting for discounting effects. The respondents were not preference-indifferent concerning the sequence of health states presented. Detailed analysis at the individual level indicated that the overall sequence effect was attributable to two groups of respondents who were sensitive to the sequence of events. One small group, referred to as "best-things-first" respondents, preferred the best years first; the other group, classified as "happy-end" respondents, preferred the reverse sequence. The majority of the respondents, however, were indifferent to the sequence.

These results suggest that 1) in valuation experiments involving the time-tradeoff method and 2) in applying valuation results to the evaluation of real-life health consequences, a varying lifetime health profile may not be regarded as simply a chain of independent separately valued and discounted QALY periods. Even elementary valuation tasks cannot safely assume ignorance of prognosis, as the additive utility independence assumption of the QALY model does not hold. The sequence effect at least supplements the conventional general time-preference effect.

Key words: sequence effect, health profiles, QALY model, additive independence assumption, time preference, time-tradeoff, utility, QALY-ples

Introduction

Pliskin *et al.* (1980) were the first to present a mathematical model for the concept of Quality-Adjusted Life Years (QALYs) that incorporated all acknowledged key elements of general medical outcomes. Miyamoto & Eraker (1985) showed that according to this model, the utility for the health outcome (Q, L) where L is the survival duration and Q is a constant health state experienced during the years of survival is defined as

$$U(Q,L) = bl^rU(q), \tag{1}$$

where b is a scaling constant and r is a risk parameter.* The risk parameter can assume any value, but will usually be $0.5 < r < 1.0$, indicating risk aversion. The value of $U(q)$ is bound to the interval (0 – 1) and depends on the contents of the health state. The duration variable l can usually be determined in a straightforward manner, but r and $U(q)$ require specific empirical investigations. This model is based on the assumptions of Expected-Utility Theory (EUT) as formulated by von Neumann & Morgenstern (1953). Based on the axioms of EUT and the empirical work of Keeney & Raiffa (1976), Pliskin *et al.* pointed out that there are three empirical conditions that should be satisfied before this QALY model represents a valid utility function for health outcomes with a constant health-status level over time.

The first condition is that the two attributes, duration and quality, be mutually independent in their contributions to the utility function. This means that the $U(q)$ assigned to any health state is the same regardless of the duration of time span, l, chosen. The second condition is referred to as constant proportional tradeoff. This condition requires that the proportion of remaining life that a person would be willing to trade off for a specified quality improvement be independent of the amount of (assumed) remaining life. In other words, if a person is willing to give up three of 12 months to improve q (suboptimal health state) to q^* (full health), this implies that he or she is also willing to give up three of 12 years for the same improvement. If both the independence assumption and the proportional tradeoff assumption hold, the model as defined by Pliskin *et al.* is valid and the utility function is defined as the constant proportional risk posture over life years. Under these conditions, the imputation of r is still required. On some occasions, r may be known, but in most circumstances this parameter has to be estimated. The computation of the risk-aversion parameter, r, is not easy, and for this reason the risk-adjusted QALY model has seldom been used in actual applications. A practical approach is to assume that the utility function over life years is linear ($r = 1$). However, this assumption introduces a third restrictive condition, i.e., risk neutrality regarding life years. Conventionally, QALY calculations in applied cost-utility analysis, adding up constant or variable durations with an assumed homogeneous q, rest on these assumptions.

The model of Pliskin *et al.* was derived for a health profile with constant quality over time, and most of the theoretical discussion of the QALY concept in the literature has concentrated

* Scalar parameters are represented in this chapter in lower-case letters and vector parameters in capital letters.

on models based on such constant health states. However, such health profiles are infrequently observed in practice. As non-chronic diseases form a major part of the disease spectrum, it is important to find ways to deal with such conditions. The additive extension of the Pliskin model assumes that separate valuations of, for instance, three different health states, q_a, q_b and q_c, can be added up to combinations of these values. Such an additive QALY measure (Mehrez & Gafni, 1989; Broome, 1993; Johannesson *et al.*, 1996) for a lifetime health profile $Q_T = [q_1, ..., q_T]$, denoted as QALY (Q_T) is:

$$U(Q_T) = \sum_{t=1}^{T} U(q_t), \qquad (2)$$

where $U(q_t)$ stands for the quality-of-life in period t. For equation 2 to be valid in situations were the health status varies over time, a fourth assumption must be made: *additive utility independence* (AUI).† This condition requires that the value of a health state in period A be independent of the value of another health state in period B. Consequently any two values for different health states can be added.

The resulting measure of outcome, the QALY, has been criticized for several theoretical reasons. First, and most important, there is little empirical evidence to show that the three Pliskin conditions and the AUI condition have been met (Loomes & McKenzie, 1989; Mehrez & Gafni, 1989; Bleichrodt & Johannesson, 1997). Second, the technical procedure of the valuation of health states (determination of qs) has been questioned for its reliability and validity. Nevertheless, empirical studies have provided evidence that carefully planned experimentation produces adequate results (van Agt *et al.*, 1994; Krabbe *et al.*, 1997), while others (Anderson, 1976; Froberg & Kane, 1989b; Mehrez & Gafni, 1993) have proposed new methods. Third, whose values should count to establish b, r and q may be an issue (Froberg & Kane, 1989a; Nord, 1992). So far, the effect of the panel's background seems small (Torrance & Feeny, 1989; Essink-Bot *et al.*, 1993). Finally, Smith (1987), Carr-Hill (1989), Loomes and Mckenzie (1989) and others have pointed to the conceptual limitations and the philosophical aspects of the QALY approach, although these issues are also pertinent to any other outcome measure.

This article focuses on the assumptions underlying the common additive QALY model. Suppose we want to know $U(A)$ and $U(B)$ of the consequences of two mutually exclusive treatments A and B. Suppose further that the consequences are homogeneous for a given treatment, and that for all patients they consist of two consecutive health states (conveniently indicated as QALY-ples) indicated by $a1$, $a2$ and $b1$, $b2$, respectively. Reliable estimates exist for the durations of all QALY-ples l_{a1}, l_{a2}, l_{b1}, l_{b2}, and all separate q_{a1}, q_{a2}, q_{b1}, q_{b2} can be adequately assessed. If we adopt the common model, for all QALY-ples the estimated time periods l are multiplied with their respective q-values, and these valued QALY-ples are

† Additive utility independence does not itself imply that risk neutrality necessarily holds. However, for QALYs to be a valid utility function for non-constant health profiles, it has to be assumed that risk neutrality with respect to life years in all health states holds and that additive utility independence holds as well.

summed (with or without discounting and assuming $r = 1$). The common model assumes that the cardinal preference of $U_A[a1, a2]$ versus $U_B[b1, b2]$ can be found by comparing $U(a1)+U(a2)$ with $U(b1)+U(b2)$. We cannot test directly the validity of this QALY calculation, as accepted empirical methods for the valuation of QALY-ple chains as a whole are not yet available. However, if we assume $a1 = b2$, $a2 = b1$, we can reduce this real-life health profile comparison to two variants of the tradeoff offered in the time-tradeoff (TTO) method. Essentially, the experiment described in this article asked participants to value a set of health states (q_i) using TTO on two occasions. The tradeoff consisted of a non-constant health profile of an excellent health state (QALY-ple *a1*) and a very bad health state (QALY-ple *a2*), with constant total duration T. By comparing the indifference value $\dot{y}$ arrived at with the sequence *[a1, a2]* with the indifference value $\ddot{y}$ arrived at with the sequence *[a2, a1]*, we were able to test for the assumed sequence effect (Gafni, 1995), as $\dot{y}$ should equal $\ddot{y}$. If the experiment confirmed the existence of a unique and distinguishable sequence effect, reflecting a principal issue involved in real-life health profiles, then a necessary assumption underlying the QALY model for non-constant health states needs to be re-evaluated.

Methods

Study: general

The empirical task was part of a large standardized experiment that included two sessions, separated by a ten-day interval. We used both individual and collective (group) response elicitations. Individual data collection was conventional; collective data collection was used to compare average individual responses with group responses for the same tradeoff tasks (results published elsewhere (Krabbe *et al.*, 1996)). For the study presented in this paper, collective responses were collected at both sessions. The participants involved in the experiment were 104 students, most of them medical students. They were seated in a lecture hall with appropriate space (> 3 m) between them.

Health states

The descriptions of the health states to be valued were constructed using the EuroQol system (Brooks, 1996). The EuroQol classification describes health status according to five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each dimension has three levels, i.e., "no problems" ("1"), "some problems" ("2"), "severe problems" ("3"). Health-state descriptions are defined by taking one level for each dimension (e.g., the best health state is represented by 11111). Theoretically this set of dimensions and levels of the EuroQol instrument allows for 243 (3^5) different health-state descriptions. Thirteen health states were selected for this study. Within each separate experiment, the descriptions of the health states to be valued were presented in a randomized order to avoid memory effects.

Elicitation of a value: the TTO operationalization

Two operationalizations of TTO were applied in the present experiment, differing only with regard to the presentation of the tradeoff of good and bad years. In both options the audience in the lecture hall was shown two slides next to each other on a big screen (see: Figure 1). The stationary health state (EuroQol health-state description) for the next ten years was projected on the left. On the right, the "tradeoff slide" was projected. Here, a bar represented ten years, divided into "good" and a "bad" periods. Numbers and marks divided the bar into 20 half-year parts. Operationalization I (first session) presented each tradeoff standard as a period of X good years followed by $10–X$ bad years; operationalization II (second session) presented the tradeoff in reverse order, starting with the bad years, Z, followed by the $10–Z$ good years. The starting tradeoff for each health state was five years in the "best imaginable health state" followed by five years in the "worst imaginable health state" (I: standard-order TTO) or the same tradeoff in reverse order (II: reverse-order TTO). Next, the tradeoff slide was varied until the point of indifference was reached for the group. The numbers of years of X and Z could be varied in steps of 0.5 years. For both options, the health state was assumed to return to its present level after ten years. If sequence effects were absent, X should equal $10–Z$.

Both methods were preceded by a similar verbal explanation of the method and a few test judgments. During the presentation, an instructor repeated the nature of the specific valuation method to avoid blurring the concepts. All the separate tasks had been pretested with other panels and a work schedule was used to ascertain equivalence of presentations, of explanations, etc.

The optimal health state, described in other studies as "perfect health" or "optimal health", was phrased as "best imaginable health state" in the present study to conform with the EuroQol format. Instead of the anchor "being dead" we used "worst imaginable health state", primarily for reasons of standardization between the valuation methods.

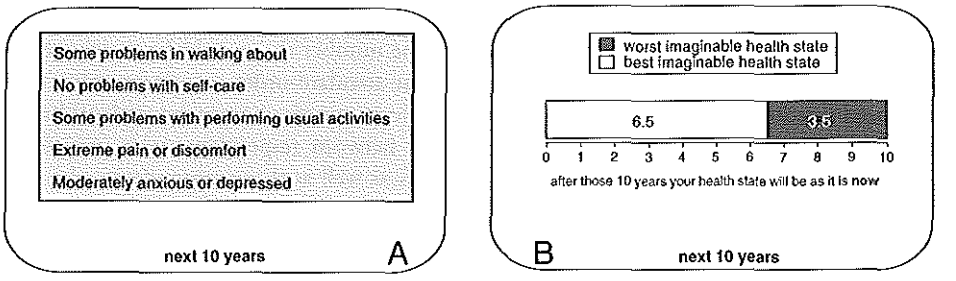


Figure 1
Example of the standard-order time-tradeoff slide presentation: the stationary health states to be valued (here EuroQol health-state description 21232) are on the left and the alternative options on the right

Individual voting system

The voting system consisted of a button-box installed at each seat in a lecture hall. The numbers of participants voting for A, the stationary state to be valued, and for B, the method-specific alternative option B, were continuously displayed on a monitor, which was accessible to the investigators only. Group indifference was defined as a situation in which the number of individual votes in favor of either option A or B was between 48% and 52%. A simple software program was developed to define consecutive bids of B, to arrive most efficiently at the indifference tradeoff. The program took advantage of the fact that the degree of dominance of one alternative provides information on the point of indifference. The voting system kept a record of all individual bids, allowing for individual analysis.

Analysis

Standard-order TTO and reverse-order TTO group indifference values were tabled with the health states arbitrarily ranked according to the standard-order TTO values.

It should be pointed out that equal tradeoffs but with different sequences will invariably yield different discounted values. For positive discount rates (d), the standard-order TTO value will increase after discounting scores. We estimated the d that, on average, produced the smallest deviation between standard-order TTO and reverse-order TTO. Thus, the most conservative estimation of the sequence effect resulted using a d-value of 5% (see: Appendix 5.1).

At a third stage, we compared the observed reverse-order TTO value with *expected* reverse-order TTO value. The expected reverse-order TTO value was based on the observed standard-order TTO value, adjusted for discounting effects, e.g., the observed standard 7.0/3.0 bid provides a discounted value of 6.1 years, which in turn equals a 2.1/7.9 reverse bid, which is rounded to an expected 2.0/8.0 reverse bid if the calibration is forced to 0.5 years.

If a systematic difference remained between observed and expected reverse-order TTO values, this would suggest a sequence effect at the group level. In that case, we intended to investigate whether subgroups or particular states would have specific sequence preferences. We identified the number of participants who consistently responded in favor of one of the two TTO modes. This number was compared to the number that would be anticipated by chance if every state valued showed a preference for standard-order or reverse-order TTO (binomial distribution, $p = 0.05$).[‡] Individual behaviors towards consecutive tradeoff bids in six health states[§] were analyzed. We distinguished between respondents who persistently overrated or underrated reverse-order TTO tradeoffs as well as those respondents who were

‡ Tested by chi-square test, assuming that due to the experimental task (reaching indifference between option A and B) and the use of the voting system, there is a considerable chance such that the final preferences for individuals does not correspond to their internal preferences.

§ This was only possible for six of the 13 health states due to the use of the voting system. For a reliable individual estimate the group consensus at the last bid has to be close to 50% preference for either A or B. For the other seven health states this was not the case.

sequence-neutral. Overraters preferred a "happy-end" scenario, whereas underraters tended to prefer to experience full health first, apart from a general discount factor.

Results

The group values elicited by the two TTO tasks for the 13 health states are presented in Table 1 (columns 2 and 4). The 9.5 value in the second column of the first row implies that their was indifference at group level between health state 11211 and living 9.5 years in "the best imaginable health state" (the state closest to the actual health state of the participants of this study) followed by 0.5 year in "the worst imaginable health state" (standard-order TTO).

Table 1
Values elicited by the time-tradeoff method (TTO; time span of ten years: 0 – 10 scale) for the 13 health states (EuroQol descriptions); responses collected by voting system (n = 100)

Health State*	Standard-Order TTO (session 1)		Reverse-Order TTO (session 2)		Expected Reverse-Order TTO Bid Based on Standard-Order TTO (after adjustment for the time-preference effect)	
	observed number of good years in tradeoff bid	5% discounted value of tradeoff bid	observed number of good years in tradeoff bid	5% discounted value of tradeoff bid	expected number of good years (5% discounting assumed)	
					expected value	expected bid
11211	9.5	7.8	9.5	7.6	9.7	9.5
21111	9.5	7.8	9.5	7.6	9.7	9.5
11112	9.5	7.8	9.5	7.6	9.7	9.5
12111	9.5	7.8	9.0	7.1	9.7	9.5
11121	9.5	7.8	9.0	7.1	9.7	9.5
11122	8.5	7.1	8.5	6.6	9.0	9.0
12212	8.0	6.8	8.5	6.6	8.7	8.5
32211	7.0	6.1	6.5	4.8	7.9	8.0
33321	5.0	4.6	5.0	3.6	6.2	6.0
21232	4.5	4.1	5.5	4.0	5.6	5.5
22323	4.0	3.7	5.0	3.6	5.1	5.0
22233	2.5	2.4	3.0	2.0	3.5	3.5
33332	0.5	0.5	1.0	0.6	0.8	1.0

* The EuroQol concept of health status consists of five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has three levels, "no problem" (1), "some problems" (2), and "severe problems" (3). Health-state scenarios are produced by the selection of one level for each dimensions (the best health state is thus represented by 11111).

Figure 2 shows graphically the group values as presented in column 4 and 7 of Table 1. For six health states the reverse-order TTO task yielded higher-than-expected values, probably due to the hypothesized sequence effect.

The individual analysis showed that a small proportion of participants could be classified as sensitive for the sequence of the events in time, preferring either best years first or the reverse, i.e., "happy-end". Four of the six health states (fifth column of Table 2) showed significant differences between the proportions of participants with a "happy-end" and "best-things-first" sequence preferences. The analysis to detect which participants responded persistently in favor of the standard-order TTO or the reverse-order TTO, showed that approximately 3% were exclusively "happy-end" respondents (three or more of the six health states of second column, Table 2). The group of "best-things-first" respondents who exclusively belonged to this category was greater, approximately 6% (three or more health states of third column, Table 2).

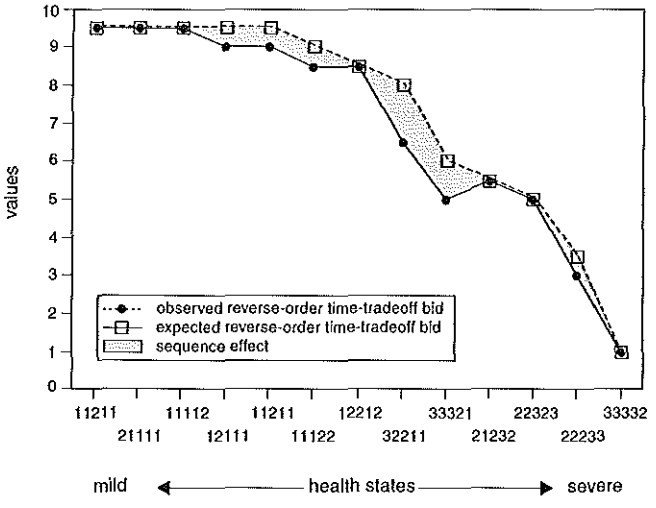


Figure 2
Observed indifference scores after adjustment for time preferences (5% discounting) for 13 health states (EuroQol classification) for the reverse-order time-tradeoff (n = 94) and the expected reverse-order time-tradeoff values

Discussion

By using the time-tradeoff (TTO) valuation method, we were able to compare the utility effects of sequences of states in a basic non-constant health profile, i.e., two-state health profiles, with the addition of a constant time period. Even this "minor" extension of the standard Pliskin model introduced persistent sequence effects, as *U[a1, a2]* did not equal

Table 2
The two shifts of preference made by the participants (sequence effect) for six health states (EuroQol classification) elicited by the standard-order time-tradeoff on session 1 and the reverse-order time-tradeoff mode on session 2 (no adjustments for time preferences)

	Shift of Preference				
Health State*	(1) Stationary Health State → (2) Tradeoff "happy end" (%)	(1) Tradeoff → (2) Stationary Health State "best things first" (%)	Indifferent "rational" (%)	Equal distributions** on Elicitation (1) and Elicitation (2)	n
12212	19	20	61	ns	98
32211***	17	9	74	p = < 0.001	96
21232	12	17	71	p = < 0.001	94
33321***	21	14	65	p = 0.02	94
22233***	19	13	68	p = 0.03	95
33332	17	22	61	ns	101
All six states	17	16	67	-	578

* Health states that proved suitable for individual analysis (at last bid about 50% of the group in favor of option A and the other 50% of the group in favor of option B).
** Chi-square test.
*** Health state with an overall "sequence effect".

U[a2, a1] even with discounting adjustment. The effect at least partially represents personal characteristics, as was shown in the individual analysis.

Internal validity of the sequence finding was supported by the fact that the effect was present over the whole range of states and not only in the intermediate states, where standard-order and reverse-order TTO values were the most divergent (see: Appendix 5.1). As consistency of the responses in general was fairly high (Krabbe *et al.*, 1996), we believe the effect reflects a true response effect. Several response effects may underlie the phenomenon. For instance, a poor health state may be more tolerable if it is perceived as a temporary hardship before finally achieving good health. Conversely, good imminent health may not seem enjoyable if there is a prospect that a substantial amount of time will be spent later in a deteriorated state that may be a prelude to death. Both time *period* effects may be estimated using a discounting approach. Alternatively, we can assume a disutility with the *event* of a deterioration or a positive utility associated with improvement. We believe that the effect of prognosis cannot be simply controlled in tradeoff devices (Sutherland *et al.*, 1982; Kassirer, 1983; Hilden *et al.*, 1992). Further investigations using more than two-state health profiles for valuation, less competent panels, a more refined scale and even more direct statements about prognosis are required.

External validity relates to the question of the extent to which the sequence effect in the TTO device can be established in real-life health profiles. The experiment started from the assumption that as valuation methods, both the TTO and the standard-gamble (SG) have to

satisfy the same conditions as the QALY concept in general. Gafni & Torrance (1984) argued that for tradeoff tasks at least three effects of the stimulus-response modeling can be distinguished, i.e., a gambling effect, a time effect, and a quantity effect. The gambling effect[‖] is intentional and is specific to the SG method. The time effect (i.e., time preference) is the effect related to the phenomenon that people generally assess acquisitions that can be gained earlier as more valuable than these acquired later, although the effect with life years seems less clear. Small but significant effects may occur in TTO (also SG) if the time spans of the constant state or the tradeoff states are changed. Time-preference parameters can be estimated from multiple-TTO tasks.[¶] The quantity effect is a context effect related to the actual number of the quantities in the task.[¥] The attitudes towards one-, five- and ten-year tradeoffs in TTO studies may appear inconsistent, as may the attitudes towards (small) probabilities in SG studies (Wakker & Stiggelbout, 1995). Sequence effects may now be added (Gafni, 1995; Richardson *et al.*, 1996). Whether all response effects have their counterparts in the values of complex health profiles requires investigation in careful well-designed experiments. Here we may need chronic stable real-life health profiles with interspaced alternative health states of fixed duration (an episodic QALY-ple) that differ only regarding the timings of the period. For instance, Lipscomb (1989) suggested such an alternative evaluation procedure that comes close to measuring a "pure time preference effect", which is truly disentangled from sequence effects. Richardson *et al.* (1996) examined the validity of the additive QALY model – an experiment without sequence effects – and also found no evidence for the additive utility independence (AUI) assumption. In an experimental study dealing with non-health assessments, it was detected that people discount outcomes differently in situations where sequences of the outcomes are involved (Loewenstein & Prelec, 1993). The recent study of Kupperman *et al.* (1997) showed sequence effects in the context of prenatal diagnostic decision making, where issues of impaired offspring were involved. However, their results may depend on the unique features of decisions associated with pregnancy, rather than decisions regarding health.

‖ The gambling effect can be thought of as synonymous with the effect that is formalized by the *r* parameter of the constant-health QALY model. Notice that the definition and interpretation of *r*, as part of the formalized QALY model, are not unequivocal. Even among (health) economists, there is considerable confusion about the meaning and interpretation of this risk-aversion parameter. We use the term "risk aversion" here solely when we are talking about the *r* parameter of the mathematical QALY model, while the concept "risk attitude" in our view stands for a much broader concept. Risk attitude covers all effects and biases related to the process of utility assessment and is not restricted to the formal QALY model. Specific effects and biases as induced by performing, for instance, the standard-gamble method, such as cognitive weighting of probabilities (prospect theory) are in our classification not captured by the *r* parameter (see also: Richardson, 1994).

¶ Due to the specific operationalization of the TTO task, the tradeoff of time, the results of the values elicited by this method are biased because of time preference, which effect is intrinsic to the TTO method. Elicited values using the TTO method therefore have to be (individually) adjusted with the same discount percentage as the discount percentage that is to estimated by the multiple TTO tasks.

¥ Here an economic notion is transformed into the domain of health psychometrics. The effect is formulated by Gafni and Torrance (1984) as: "the quantity effect asserts that, all else equal, one's preference for the next (marginal) unit of time in a given health state depends on the length of time already spent in that state". We may question whether, for the situation of health, the time preference effect and the quantity effect are not essentially the same.

Applying the simplest valuation devices (e.g., rating scales), it may be feasible to compare deliberately chosen health profiles (QALY-ple branches without ramifications) to specify the direction and the magnitude of the effect. The healthy-years equivalents (HYE) approach as proposed by Mehrez and Gafni (1989) already moves in the direction of valuation of defined real-life health profiles:

$$U(HYEs, q^*) = U(Q), \tag{3}$$

where Q is a vector of the health states in each year, i.e., $Q = (q_1, q_2, ..., q_i)$ and HYEs are the numbers of years in full health (q^*) that can be considered equivalent to the health profile. As Weinstein & Pliskin (1996) has pointed out, HYEs permit more flexibility in assigning numbers of QALYs to health profiles, as the AUI assumption is no longer required. However, the demands made by the assessment task for HYEs are undoubtedly greater than those for QALYs, and so the feasibility of the HYE has to be proved. Finally, our results urge the need to investigate the gambling effect of SG with its obvious counterpart in real-life health profiles, viz, ramifications of a particular course of disease (a QALY-ple tree). Suggestions are already made (Johannesson, 1995) to use the HYE approach for health profiles framed in terms of uncertainty. We question whether the effect of probability patterns present in the health profile on $U(Q)_T$ is necessarily the same as the one found in the rather simplified SG values of $U(q_t)$. The gambling effect may be estimated similarly to the sequence effect (by means of a r_{SG} or a specified disutility) from multiple SG tasks. Subsequent comparison with the effect found in real-life health profile valuation should demonstrate whether this r relates to the r referring to the value of health profiles as a whole.

Some limitations associated with this study should be addressed. First, memory effects could have biased the results. However, if memory effects had existed, they would have deflated, rather then exaggerated, the effect of the sequence. Moreover, the interval between the two sessions, the similarity of the unlabeled stimuli, and the randomization of the presentation orders almost rule out memory effects (this was further supported by feedback from the respondents). Furthermore, our intentional choice for a within-subjects design allowed us to use statistics with far greater power compared with a between-subjects design (e.g., 50 subjects standard-order TTO, 50 subjects reverse-order TTO). Another aspect to be considered was the use of "best imaginable health state" and "worst imaginable health state" as TTO anchor points (EuroQol format). The use of a "vital" health state instead of "being dead" is necessary to create symmetric standard-order/reverse-order TTO tasks. More generally, the use of "being dead" in cases of valuation of *"partial"* durations is complex, e.g., the issue of how to value a period that follows another one judged to be worse than "being dead". From a methodologic point of view, the TTO method allows any set of anchor points to be chosen, as is the case with all tradeoff response techniques (Llewellyn-Thomas *et al.*, 1982; Torrance, 1986). The elicited values, of course, need recalibration to be comparable with other studies. Based on Expected-Utility Theory we recently showed (Krabbe *et al.*, 1996) that the values of our experiments are a linear combination of "real" utilities.

As a practical consequence of this experiment, we consider that although the assumptions underpinning AUI are not upheld, this does not necessarily mean that the more general

extended QALY model should be rejected. First, the size of the effect seems reasonably small, and most cost-effectiveness results will not be affected, as effect differences between treatment options usually occur in early follow-up. From a more theoretical point of view, it is important to know whether the sequence effect is a general effect expressed by the large majority of persons to some degree, but not for all health cascades, or whether it represents an unevenly distributed personal characteristic. In the first case, specific experimental valuation tasks could be performed to elicit a parameter for the sequence effect either as time-utility modifier or as a zero-time health-status-utility modifier. For example, if we define $U(L)$ if followed by a same or better QALY-ple as $U(L) = 1.0$ and $U(L) = 0.5$ otherwise, we take account for the sequence effect by a time-utility modification. A similar approach is valid for the gambling effect, if it exists. If, however, the sequence effect is uniquely attributable to individual preferences, the validity of the extended QALY model as a general population's proxy becomes endangered.

Two strategies are available to overcome the shortcomings of the present utilization of QALYs. First, we may extend and refine the simple, though robust QALY model. This strategy undoubtedly requires a considerable effort in experimental studies and an extension of Pliskin's model. As a consequence of such a generalized model, we are inevitably forced to develop for each specific parameter ("valuation of health" simply, time preference, risk aversion, sequence effect, ...) an appropriate measurement procedure. Second, instead of the modeling strategy, we may leave conventional measurement procedures and look for new and better appraisal settings that confront people with realistic health situations and include all the relevant elements in the (composite) measurement procedure itself. To make progress in this area and to arrive at adequate and genuine valuations of health, more work should undoubtedly be undertaken.

Acknowledgment

The authors thank the Advisory Board for the Health Research Promotion Program (Adviesgroep SGO) for financial support of the Research Program "Standardization in Medical Technology Assessment" and M.L. Essink-Bot and H. van Agt for their contributions to the field work. They also acknowledge the useful comments and suggestions made by three anonymous referees and by the editor.

Appendix 5.1

Figure A1 presents the values of ten year tradeoffs after discounting with discount rates (d) of 1% and 5%. For example, the 5% discounted "5 years good"/"5 years bad" bid equals a discounted value of 4.55 (standard-order TTO) versus 3.56 (reverse-order TTO), with the assumption $U(q_{\text{worst imaginable}}) = 0$. We estimated the d that 1) on average produced the smallest deviation between standard-order TTO and reverse-order TTO and 2) yielded equal values for the best and for the worst health state(s). This second condition is required by definition, as can be seen in the figure.

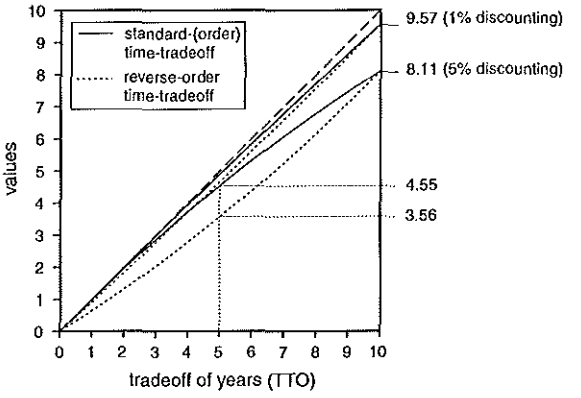


Figure A1
Representation of the bias introduced by the use of the standard (order) time-tradeoff method (time horizon ten years) based on an overall time preference (1% and 5% discount rates are depicted)

Chapter 6

The Effect of Adding a Cognitive Dimension to the EuroQol Multiattribute Health-Status Classification System

Paul F.M. Krabbe, Marlies E.A. Stouthard, Marie-Louise Essink-Bot, Gouke J. Bonsel
(submitted for publication)

Abstract

A methodological study was conducted to examine the effect of extending a frequently used simple multiattribute health-status classification system by adding a cognitive dimension.

The EQ-5D questionnaire is a generic instrument to value health, developed by the EuroQol Group. The EQ-5D defines health according to five dimensions: mobility, self-care, usual activities, pain/discomfort, anxiety/depression. We defined 18 different health states according to the standard EQ-5D classification. A second set of health states was constructed similar to the first, except for the addition of a cognitive dimension (EQ-5D+C). Valuations of both sets of health states were statistically analyzed to detect the effect of the additional dimension.

The cognitive dimension generated systematically different values compared to the standard EQ-5D version, whereas the content validity improved. Both systems evoked equally reliable values. Analyses showed that a simple additive model to predict summary values for health states was not optimal for both systems.

Although there is a current lack of consensus regarding the domains that are selected to represent health status, this study has shown the importance of considering the inclusion of a cognitive domain.

Key words: multiattribute health-status classification systems, cognitive dimension, validity, reliability, methodology, EuroQol

Introduction

During the last twenty five years, considerable effort has been invested in the development of summary measures which allow for valid comparison of health status (health-related quality-of-life) across different diseases. Implementation of such preference-weighted measures ranges from the assessment of an individual patient's preference for treatment outcomes, as occurs in clinical decision making, to planning and monitoring health programs conducted by health policy makers. The goal is to capture a single numerical value that represents the relative desirability of a particular health state.

Basically, two different valuation strategies can be distinguished. The early pioneering work on the measurement of values for health states was based on *naturalistic* or *holistic* designs of the outcomes (see: Torrance, 1986; Froberg & Kane, 1989a). If there are only a few final health-state outcomes of interest, which is usually the case in cost-utility analysis and in individual clinical decision analysis, each state can be described separately. Subsequently, the values (utilities) for such states are arrived at by the valuation of naturalistic descriptions (sometimes called vignettes or scenarios), such as "blindness" or "angina pectoris with pain at rest", using one of the standard valuation methods. This design is obviously not suitable if a considerable number of different health-state values have to be obtained as in multiform diseases or, more generally, in population studies.

In later studies, *decomposed* designs were applied (Froberg & Kane, 1989a; Gold, 1996; Drummond *et al.*, 1997). Prior to valuation, each clinical health state is mapped on a predefined general classification. By doing so, the total number of states to be valued is determined by the possible permutations of the classification. If the number is small, it may even be feasible to value them all. If the number of states is moderately large, a well-chosen subset can be valued empirically, while the values for the remaining states can be estimated by some formal method (usually regression modeling). To allow for efficient statistical modeling, the classification should consist of a set of multiple dimensions (formally called attributes), with multiple, hierarchical levels per attribute.

Instruments that support a decomposed design are the Quality of Well-Being scale (QWB) (Kaplan *et al.*, 1976; Kaplan & Anderson, 1988), the Health State Classification System (HSCS; renamed as HUI-1) (Torrance *et al.*, 1982) and its successors the Health Utility Index Mark 2 and 3 (HUI-2 and HUI-3) (Feeny *et al.*, 1994; Torrance *et al.*, 1995; Torrance *et al.*, 1996), and the EuroQol (EQ-5D) instrument (EuroQol Group, 1990; Brooks, 1996). The structure of attributes and levels is different for these four health classification systems, namely: QWB, three attributes, 3/5 levels and 27 specific symptoms; HUI-1, four attributes and 4/8 levels; EuroQol, five attributes and 3 levels; HUI-2, seven attributes and 3/5 levels; HUI-3, eight attributes and 5/6 levels.

Although all these classifications are multiattribute systems (Keeney & Raiffa, 1976; Fischer, 1979; Torrance *et al.*, 1982), the respective authors have proposed different strategies for the estimation of values for those health states that were not valued empirically (Feeny *et al.*, 1992; van Hout & McDonnell, 1993; Abdalla & Russell, 1995; Ohinma *et al.*, 1996; Torrance *et al.*, 1996; Dolan, 1997).

It is generally agreed that attributes should be derived from a comprehensive concept of health status, covering at least the physical, psychological and social domains. Nevertheless, the different classification systems are not yet comparable and empirical research comparing their health constructs has hardly been conducted. Intensive discussions about the conceptualization of the classification of "health" were held at the inception of the EuroQol Group, which culminated in the choice for a "simple" classification intended to be suitable for measuring across the full health spectrum (Brooks, 1996). As a consequence, the EuroQol Group recognized that the EQ-5D classification could not claim to be a comprehensive measure of health status. In this paper we assess the comprehensiveness of the concept of health status as defined by the EQ-5D classification (Essink-Bot *et al.*, 1993; van Agt *et al.*, 1994; O'Hanlon *et al.*, 1994; Selai & Rosser, 1995) in a Dutch study to obtain disease-specific disability weights for a large number of diseases (Stouthard *et al.*, 1997). This study was an adapted Dutch replication of the Global Burden of Disease study commissioned by the World Bank (Murray & Lopez, 1996). The consulted national expert group in the Dutch study questioned two characteristics of the EQ-5D classification. In view of the major public health impact of cognitive functioning due to, in particular, dementia and mental retardation, the absence of a "cognitive functioning" attribute in the standard EQ-5D format was regarded as a major limitation. The same group also pointed out the lack of calibration potential as all EQ-5D attributes are divided into only three levels. An extension of the standard EQ-5D version by adding a "cognitive functioning" attribute was proposed (EQ-5D+C). The three-level system however was left unaltered (Bonsel & van Agt, 1995; Selai, 1996). In the present study, valuations elicited from the resulting EQ-5D+C descriptions were compared empirically with parallel standard EQ-5D descriptions.

We assumed that the introduction of an additional attribute to a classification system would systematically alter the valuations in the following ways.

- I) By adding the Cognition attribute, the comprehensiveness of the concept of health status will improve (content validity).
- II) This improvement however will be countered by increased variability in the responses, due to the increased complexity of the valuation task (reliability).

We were uncertain whether the advantage of I would outweigh the disadvantage of II. In addition, we were interested to know whether a simple additive model (main effects only), instead of a factorial model (main effects and interaction effects), would be sufficient for both the five- and six-attribute classification to estimate a value function that was sufficiently precise to predict non-valued health states.

Material and Methods

The classification system

For the description of health states, we used the standard EQ-5D classification system developed by the EuroQol Group (Brooks, 1996). The EQ-5D classification describes health status according to five attributes: mobility, self-care, usual activities, pain/discomfort and

anxiety/depression. Each attribute has three levels, i.e., "no problems" ("1"), "some problems" ("2"), "severe problems" ("3"). Health-state descriptions are constructed by taking one level for each attribute (e.g., the best health state is represented by 11111). Theoretically this set of attributes and levels of the EQ-5D instrument allows for 243 (3^5) different health-state descriptions. For the present study, we constructed an extended version (EQ-5D+C) of the standard EQ-5D which included "cognitive functioning" as an additional attribute, with a similar operationalization of three levels: (1) no impairment of cognitive functioning, (2) some impairment of cognitive functioning, (3) severe impairment of cognitive functioning. This non-standard EQ-5D+C classification would allow for 729 (3^6) different health-state descriptions. As for the conventional classification, where the attribute "usual activities" is clarified by mentioning specific areas of interest "work, study, housework, family or leisure activities", cues for the Cognition attribute were also presented as: "memory, concentration, coherence, IQ".

Selection of the health states to be valued

From a population study conducted in 1993 (Essink-Bot *et al.*, 1993) we selected 14 out of 25 health states described by the EQ-5D version. Only one selection criterion was used: a representative coverage of the scale, that is an even mix of good, moderate and bad health states according to the assigned values. The 6th attribute was added to all these states, choosing the level randomly, occasionally avoiding a level too unlikely in combination with the other five levels. The best (11111[1]) and the worst health state (33333[3]) were also included. To these 16 states, two other EQ-5D health states, which formed no part of the 1993 study, were added to construct a parallel set of EQ-5D+C health states with assumed significant effect of the attribute "cognitive functioning". Finally, the EQ-5D+C set comprised two extra health states differing only one level in Cognition level from an already selected EQ-5D+C state, enabling to study the single effect of varying the level of the Cognition attribute, with a background of different severity (212321 vs. 212322, 323331 vs. 323332). In all, 18 health states were constructed for the EQ-5D version and 20 states (18 counterparts of the EQ-5D version) for the EQ-5D+C version.

Valuations

Evaluation of the health-state descriptions of the EQ-5D and the EQ-5D+C was performed by the standard EuroQol method. The EuroQol rating scale is presented as a vertical thermometer with a scale from 0 to 100. The anchors are labelled "best imaginable health state" at the top of the thermometer (100) and the "worst imaginable health state" at the bottom (0). In the present study, all the descriptions were presented in boxes on two pages of the questionnaire. On one page, five boxes were placed on either side of a vertically placed thermometer (rating scale), while the remaining eight (EQ-5D) or ten (EQ-5D+C) boxes were placed on the facing page. The respondents' task is to locate the health-state descriptions on the scale between the two anchors by drawing a line, in such a way that the intervals between the scale positions of the descriptions correspond to the differences in preference as perceived by the participant (Figure 1).

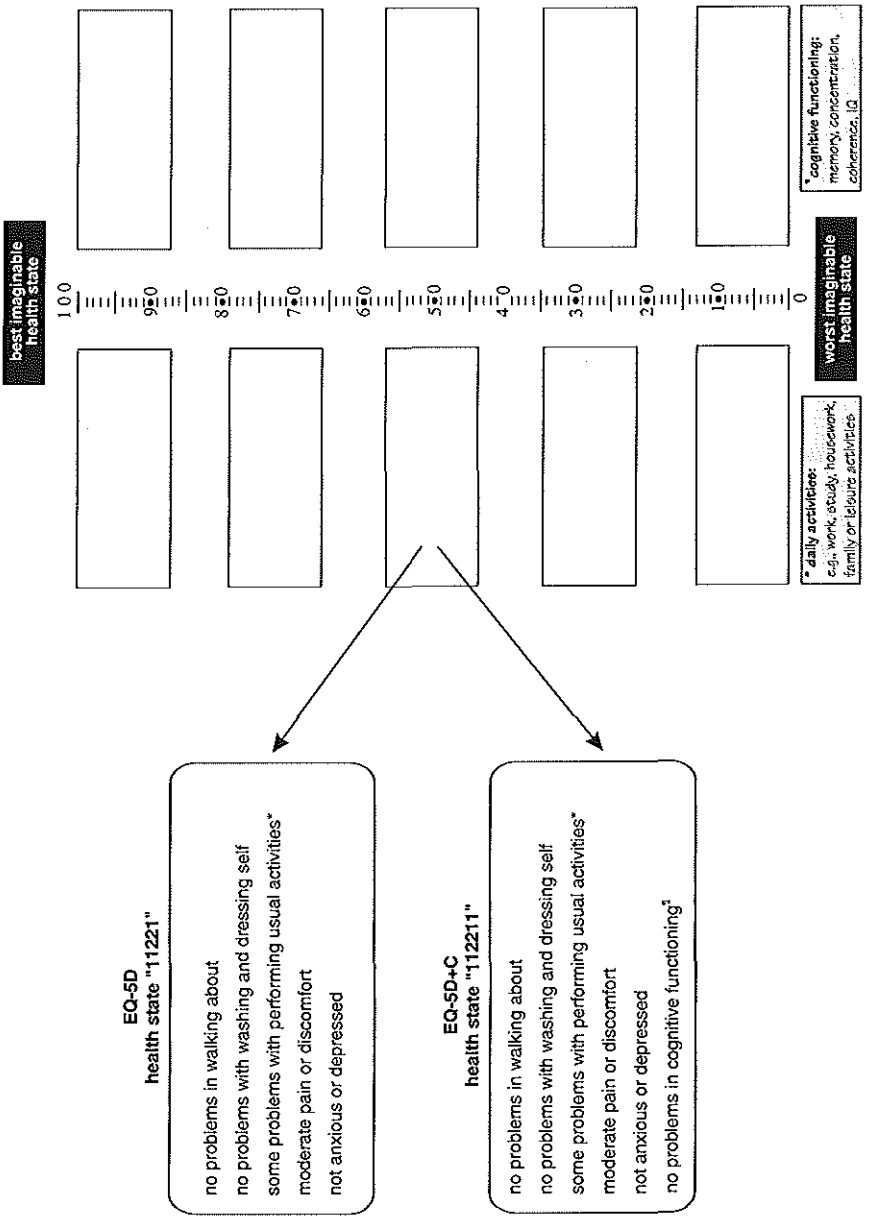


Figure 1
Example of the valuation task for the standard EQ-5D and the EQ-5D+C

The respondents were instructed that the duration of the health states was one year. What happened after this period was stated to be unknown.

Study design

Potential respondents in this study were all Dutch faculty members on three different locations (the two departments and the institute were the authors were appointed), none of them engaged in the current study. All 185 faculty members (scientific staff members and management members) received a questionnaire in their mailbox with a covering letter. In this letter the addressed were requested to participate in the study after explanation of the purpose of the study. Reminders were not issued and no fees were paid to participants for their contribution. Questionnaires containing the EQ-5D and EQ-5D+C versions of the classification system were randomly distributed.

Analysis

Background characteristics of the two groups of participants, i.e., those who had produced valuations for the EQ-5D and the EQ-5D+C classification system respectively, were summarized. Means and standard deviations were computed for all the valued health state descriptions. Statistical testing within pairs (EQ-5D vs. EQ-5D+C) of descriptions was performed by t-tests. Subsequently, for some specifically constructed health states, e.g., replications and health states with level variations, we were able to test for significant differences within each version (paired t-tests).

Two less common statistical techniques were used, which are specifically suitable in this context. Multiple Classification Analysis (MCA) was used to examine the *relative* importance of the attributes. The MCA technique is closely related to analysis of variance* and is almost identical to multiple regression analysis with coding ("dummy") variables (Andrews *et al.*, 1973; SPSS inc., 1997). Opposed to these conventional analytical techniques, MCA is based on an iteration algorithm instead of matrix inversion for estimating the coefficients. Due to this iteration-based algorithm MCA is more flexible to deal with correlated predictors (attributes). The main advantage of MCA is that coefficients for all levels are obtained and expressed as deviations from the overall mean, which enables the researchers to assess the relative magnitude of the effects of the levels for each attribute separately. A key feature of the technique is its ability to show the effect of each attribute (e.g., predictor) on the valuations of the health states (e.g., the dependent variable) both before (unadjusted) and after (adjusted) taking into account the effect of all other attributes. However, adjusted deviation scores can only be safely interpreted if there are no substantial statistical interactions among the attributes.† These two association coefficients represent the relative contribution of each separate attribute as a whole, with (beta) and without (eta) adjustment for dependency

* Option in SPSS: simple factorial analysis of variance.

† Interactions between attributes must not be confused with correlation between attributes. Correlation deals with the question of whether attributes as a whole are independent of each other, whereas interactions deal with combination of levels of different attributes that affect each other.

between the attributes respectively. Eta coefficients should be interpreted as a measure of the ability of the predictor to explain variation in the dependent variable (health states) independent of the effect (contribution) of all other predictors. Beta coefficients should be interpreted as a measure of the ability of the predictor to explain variation in the dependent variable (health states) after adjusting for the effects of all other predictors.‡ For the betas, this is not done in terms of percentage of variance explained (estimation not based on standard regression modeling). However, the rank order of these betas indicates the *relative* importance of the various predictors (attributes).

Finally, analysis based on generalizability theory (G-theory), an analysis of variance technique, was performed to estimate the contribution of the different "variance components" (health states, respondents). In G-theory, the proportion of variance of the component "health states" reflects the amount of internal consistency of the respondents' responses to the different health states and may be interpreted as the reliability of the classification system used (Streiner & Norman, 1995). Hence, MCA was used to test validity of the content of the EQ-5D and EQ-5D+C classification systems, whereas G-theory compared the reliability of these two classification systems.

Analyses were not performed on calculated, aggregated data, but on the complete data set. Analysis of averaged data increases the total proportion of variance explained, because the individual variation is eliminated. However, taking individual response into consideration is vital for making comparisons with other studies and is also required to make adequate adjustments for individual background variables and individual response behavior (Ohinma *et al.*, 1996; Dolan, 1997). No prior transformation of the individual data was considered, while, apart from interpretation problems, previous studies did not substantially benefit from transformation of the dependent variable (Abdalla & Russell, 1995; Krabbe *et al.*, 1995).

Results

Response and background characteristics

A total of 87 valid questionnaires were returned, 39 comprising the EQ-5D version and 48 comprising the EQ-5D+C version, i.e., a response rate of 47%. The rather low response rate was partially explained by the fact that about 60 faculty members were on holiday at the time the study was scheduled (July). Of the total group of respondents, 42% were men. The mean age of the respondents was 33 years with a standard deviation of eight years. The characteristics of both groups of respondents were approximately the same, including a good state of health (see: Table 1). No complaints or difficulties were noticed about the stimulus

‡ The beta measure is analogous tot the standardized (partial) regression coefficient, i.e., the regression coefficient multiplied by the standard deviation of the predictor (attribute) and divided by the standard deviation of the dependent variable (health state value), so that the result is a measure of the number of standard deviation units the dependent variable moves when the exploratory variable changes by one standard deviation. Another name for such a coefficient is the "partial beta coefficient".

and the valuation task. Some responders questioned the soundness of valuing health states in general.

Descriptive statistics

In Table 2, the mean valuations and standard deviations for the health states are shown, together with the results of t-tests for differences between EQ-5D and EQ-5D+C valuations.

Significant differences between EQ-5D and EQ-5D+C were found for health state 11211[2], where a "moderate" Cognition level was added to a "good" state; 12311[3] and 13311[3], where in both cases a "bad" Cognition level was added to a "moderate" state; 12212[1], where a "good" sixth level was added to a "moderate" state; 32333[1], where a "good" Cognition level was added to a "bad" state; and 33333[3], where a "bad" Cognition level was added to a "bad" state. The latter four differences did not reach distinct statistical significance (p-values ranges: 0.01 – 0.05).

Adding a "good" Cognition level to a bad health state resulted not in a meaningful improvement considering the valuation of these states (32211[1], 21232[1], 33321[1], 32333[1]). These states were valued somewhat better, but only for health state 32333[1] the difference was statistically significant. Adding a "bad" Cognition level to a "bad" health state made the valuation far worse (Table 2). In general, adding a non-neutral sixth level to a EQ-5D health state changed the valuation. In Figure 2, the largest shifts in valuations between EQ-5D and EQ-5D+C were observed in health states where a sixth level was added that was different from the other five levels.

Generally, the standard deviations for the valuations elicited by the EQ-5D+C version were slightly greater than those for the EQ-5D version.

Table 1

Background characteristics of the 87 respondents (means with standard deviations between parentheses or percentages) who assessed the standard EuroQol classification system (EQ-5D) and an extended version (EQ-5D+C)

Background Characteristics	EQ-5D (n = 39)	EQ-5D+C (n = 48)
Age (years)	33.0 (7.8)	32.9 (7.9)
Sex	50%	36%
male	50%	64%
female		
Valuation own health status*	87.0 (12.0)	89.6 (10.8)
Experience with severe illness		
respondent self	0%	6%
relatives of respondent	65%	67%
professionally	19%	23%

* Valuation range from zero (worst imaginable health state) to 100 (best imaginable health state).

Table 2
Means and standard deviations of valuations of health states based on the EQ-5D version* and the EQ-5D+C version**

EuroQol Health States	EQ-5D (n = 39)		EQ-5D vs. EQ-5D+C	EQ-5D+C (n = 48)	
	Mean	SD	p-Values	Mean	SD
11111(1)	95.9	5.7	ns	95.8	13.4
11111(1) replication	95.7	5.7	ns	95.2	13.7
11121(1)	78.0	11.0	ns	77.8	14.7
11211(2)	77.0	11.5	< 0.01	55.4	20.3
11112(1)	66.0	16.7	ns	68.2	17.7
21211(1)	63.7	12.4	ns	64.6	18.6
11221(1)	63.5	11.9	ns	65.4	17.6
11122(1)	57.0	17.2	ns	60.2	18.8
12311(3)	56.5	13.4	< 0.01	32.7	20.2
21212(1)	54.4	13.2	ns	58.7	18.1
12212(1)	53.3	15.1	0.04	61.3	19.1
13311(3)	49.4	13.7	< 0.01	28.8	18.2
32211(1)	44.1	17.4	ns	47.9	17.4
21232(1) A'	34.6	11.4	ns	39.3	17.7
21232(2) A"	-	-	-	29.4	16.3
33321(1)	31.8	15.3	ns	35.5	17.5
33333(3) replication	13.1	16.5	0.03	6.4	10.3
32333(1) B'	12.2	12.9	0.02	19.1	13.7
32333(2) B"	-	-	-	15.6	13.8
33333(3)	10.2	14.8	0.03	4.6	8.2

* The standard EQ-5D classification system of health status consists of five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has three levels, "no problem" (1), "some problems" (2), and "severe problems" (3). Health-state scenarios are produced by the selection of one level for each dimension (the best health state is thus represented by 11111).

** Similar to the standard EQ-5D version except for the additional dimension "cognitive functioning" (memory, concentration, coherence, IQ).

Equivalence of valuation

In each version, two health states were presented twice (11111[1], 33333[3]). Valuations were compared for both pairs within each version as a reliability test. Paired t-tests showed no differences in valuations for health state 11111[1] for either version, and a small but significant difference in valuation for health state 333333 for the EQ-5D+C version only.

In the EQ-5D+C version, two health states were presented twice changing only the level of the sixth, Cognition attribute (21232*1* and 21232*2*, 32333*1* and 32333*2*). Equivalence of the valuations was tested with a paired t-test. For both pairs, valuations differed significantly. The effect of level variation was as expected: the "better" level resulted in a higher valuation for that health state (resp. 39.3 vs. 29.4, 19.1 vs. 15.6; see: Table 2 and 3), and the valuation for the bad health state in the second pair became lower if the Cognition level was changed from 1 to level 2.

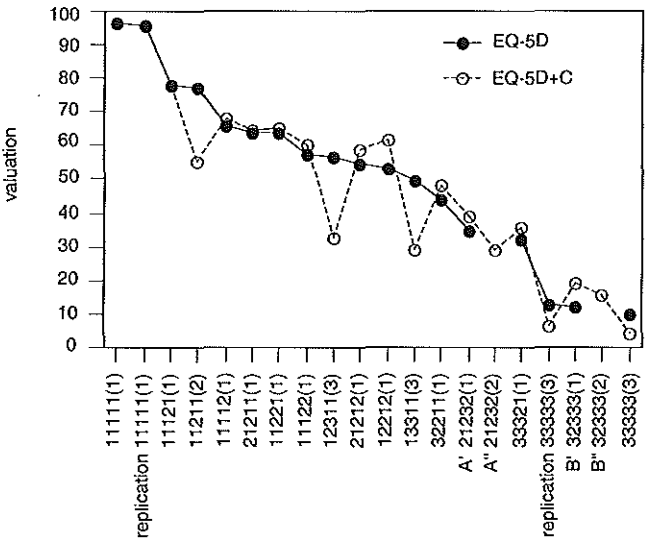


Figure 2
Mean valuations of the EQ-5D and the EQ-5D+C
(states arbitrarily ranked following mean values standard EQ-5D version)

Table 3
Reliability of anchor states (EQ-5D versus EQ-5D+C)
and level effect of the Cognition attribute for two non-optimal health states

Pairs of EuroQol Health States	EuroQol version	Difference	p-value
Reliability			
11111 vs. 11111	EQ-5D	0.2	ns
111111 vs. 111111	EQ-5D+C	0.6	ns
33333 vs. 33333	EQ-5D	2.9	ns
333333 vs. 333333	EQ-5D+C	1.8	0.03
Level Effect			
212321 vs. 212322	EQ-5D+C	9.9	< 0.01
323331 vs 323332	EQ-5D+C	3.5	0.02

Contribution of the attributes and their levels

For each level of the attributes MCA presents the unadjusted deviation scores from the overall mean. If the attributes of the EQ-5D and EQ-5D+C EuroQol version were statistically independent of each other (implying that the five resp. six attributes each represent a unique aspect of the general health-status concept), a simple additive multiattribute model could be constructed based on these deviation scores. Table 4 shows that if the deviation scores are added up an invalid sum-score may results at the extremes: for instance, if the EQ-5D health state "11111" is reconstructed by the formula: overall mean + deviation for the corresponding levels of each attribute, a score beyond 100 results (i.e., 53.2 + 16.2 + 15.5 + 25.5 + 12.6 + 12.4 = 135.4). If we estimated a model with adjustment for mutual (whole domain) dependencies, reversals occurred, where "reversal" implies that level 3 (or 2) is preferred above level 2 (or 1).

Table 4
The relative contribution of the attributes and their three levels to the valuation of health states for the standard EQ-5D and the EQ-5D+C

Attributes and Levels	EQ-5D (grand mean = 53.2)	EQ-5D+C (grand mean = 48.2)
Mobility		
1	16.2	16.0
2	-2.6	-0.2
3	-30.8	-26.8
Self-Care		
1	15.5	16.4
2	-11.6	-12.4
3	-27.0	-29.4
Usual Activities		
1	25.5	31.3
2	2.5	4.6
3	-24.3	-27.8
Pain/Discomfort		
1	12.6	12.9
2	4.4	11.3
3	-35.8	-29.2
Anxiety/Depression		
1	12.4	11.8
2	-0.3	4.7
3	-41.3	-36.9
Cognition		
1	-	12.7
2	-	-14.5
3	-	-30.2

* The effect of each attribute (e.g., predictor) on the valuations of the health states (e.g., the dependent variable) without taking into account the effect of all other attributes.

Table 4 also shows that the weights for the levels of the five conventional attributes were approximately equal, whereas the weights for level 2 were generally closer to level 1 than to level 3.

We determined the explanatory power of the attributes with MCA. The eta coefficients show (Table 5) that for both EuroQol versions the crude contribution of each single attribute was more or less the same. A notably different result was observed for the beta coefficients (Table 6). For both EuroQol versions, substantially lower contributions were noticed for most attributes. In particular, the "Self-Care" attribute showed a much lower contribution to the total assessment after correcting for dependency with the other attributes (EQ-5D: unadjusted 0.63, adjusted 0.18; EQ-5D+C: unadjusted 0.62, adjusted 0.12). For the EQ-5D as well as for the EQ-5D+C "Usual Activities" seemed the most important attribute on the EuroQol health-state classification. On the EQ-5D+C, the attribute "Cognition" was the second most important attribute.

Finally, we performed a G-study. The percentage of variance uniquely attributable to the

Table 5

The contribution of the attributes (eta coefficients) for the standard EQ-5D and the EQ-5D+C to the valuation of health states unadjusted for interdependency (expressed as the proportion of variance explained by the attribute in the valuation of the health states)

Attributes	EQ-5D (n = 39)	EQ-5D+C (n = 48)
Mobility	0.71	0.60
Self-Care	0.63	0.62
Usual Activities	0.69	0.74
Pain/Discomfort	0.68	0.62
Anxiety/Depression	0.68	0.60
Cognition	-	0.60

Table 6

The relative contribution of the attributes (beta*) for the standard EQ-5D and the EQ-5D+C to the valuation of health states adjusted for interdependency

Attributes	EQ-5D (n = 39)	EQ-5D+C (n = 48)
Mobility	0.26	0.33
Self-Care	0.18	0.12
Usual Activities	0.34	0.44
Pain/Discomfort	0.25	0.19
Anxiety/Depression	0.28	0.26
Cognition	-	0.36

* A beta is a standardized regression coefficient. It can not be interpreted in terms of percentage of variance explained.

valuation of the health states was 78% for the EQ-5D and 74% for the EQ-5D+C. These percentages of variance may be interpreted as standard reliability coefficients (Streiner & Norman, 1995) of the two classification systems, 0.78 and 0.74 respectively. The percentages of variance uniquely attributable to systematic differences among the participants was 7% for the EQ-5D and 13% for the EQ-5D+C. The remaining variance, 15% for EQ-5D and 13% for EQ-5D+C, reflects the interaction between the components "health states" and "participants" and a random error component.

Conclusions and Discussion

Several approaches exist for producing summary values for overall health which allow for combined mortality/morbidity measures such as the QALY. All rely on the empirical valuation of one or more health-state descriptions using a standardized classification system. Evidence which shows unfavorable performance aspects of an existing health-status classification system may call for additional experimentation. In this exploratory study, we investigated an often proposed improvement of the frequently used EQ-5D classification, viz., the adding of a Cognition attribute. It appeared that "good" health states were affected more by the addition of impaired cognition (e.g., 11211 vs. 11211*2* or 11211*3*). The reverse however was not true: bad health states were much less affected by the introduction of an additional attribute representing good cognitive functioning. Second, the separate contribution of each of the six attributes examined in this study was almost the same. However, after taking into account the interdependency among the attributes, we observed significant differences in the contribution of the five attributes (EQ-5D) and the six attributes (EQ-5D+C) respectively. Third, we found evidence that the simple additive model for both the EQ-5D and the version with the Cognition attribute was less adequate compared to an equally straightforward full factorial model (main effects + all possible interactions) or a restricted factorial model (e.g., main effects + first-order interaction effects). We therefore conclude that the EQ-5D and EQ-5D+C are not only characterized by overall dependencies between attributes (correlations) but also by dependencies between specific levels of different attributes (interactions). It is likely that these two types of dependencies can be expected in other classification systems as well.

By extending the standard EQ-5D system to include a cognition attribute, the concept of health status becomes more comprehensive. Statistically, Cognition is second in importance to the Usual Activities attribute and contributes to an increase in content validity. Possibly, the importance of Cognition was a direct result of the respondents in this study, who as faculty members could be expected to attach a high value to the absence of cognitive problems due to their academic background. Given the ultimately normative nature of concepts, there was no formal approach to determine the "degree" of content validity, but in our view the content improvement certainly outweighed the slight decrease in reliability. As already mentioned in 1976 by Kaplan *et al.*, a health-status concept must be constructed on the basis of substantive theory, including a conceptual check by experienced users of the instruments derived from the theory. The same holds for the more commonly used descriptive health-status (health profile)

measures (Essink-Bot *et al.*, 1997; Staquet *et al.*, 1998). Supported by theoretical and practical arguments, we considered the inclusion of the Cognition attribute. From the viewpoint of statistical and empirical efficiency, we should be aware that separate attributes do not guarantee separate effects. The use of more attributes increases the probability of empirical correlation. A second drawback of increasing the number of attributes (information) is a parallel increase in the variability of the responses, and subsequent decrease in the reliability (Payne, 1982). In our study the obvious tradeoff between the validity and comprehensiveness of the health-status concept on the one hand, and the reliability of its assessments on the other, clearly supported an extension of the EuroQol classification. We do not expect however that a further extension of domains will provide notable improvement, but refinement *within* attributes (4 or 5 levels) still needs exploration.

In this exploratory study, we were not able to test all the experimental factors, in particular the specific contribution of the many possible interaction terms and the role of personal background characteristics in our homogeneous group of respondents. Lack of representativeness of this panel precluded us from treating the weights for the different attributes as general population weights. We expect similar results in a lay population in view of previous EuroQol surveys employing different panels (Essink-Bot *et al.*, 1993; van Agt *et al.*, 1994). However, an important issue yet to be resolved is the generalizability across different European countries. The universality of the EQ-5D has been always an important claim. The reliability of 0.78 found in this study is similar to the reliability of 0.77 reported in a similar study among 104 students (Krabbe *et al.*, 1997).

This study also presented evidence that the simple additive model without interaction terms is not optimal for either the standard EuroQol classification system or the EQ-5D+C variant (Krabbe *et al.*, 1995). Interaction effects were apparently present in the EQ-5D and EQ-5D+C. The additive model may nevertheless be used as a robust device for predicting average health-state values. More detailed and controlled investigations focusing on the type of model while keeping other factors constant is required.

In this study, we developed an approach to judge the extension of a multiattribute classification system. The approach is also suitable for studying other changes, such as disaggregation of the levels. The data support our experience in application studies (clinical, public health), where experienced physicians and health administrators frequently pointed out the omission of the Cognition domain as an independent attribute. Important for the further dissemination of multiattribute health-status classification systems are systematic studies focused on the conceptualization of the associated health-status constructs, and additionally, head-to-head empirical comparison of the different classification systems.

Acknowledgment

The authors wish to thank the other members of the project "Disability Weights for Diseases in the Netherlands", of which this study forms a part, and J.J. van Busschbach for his contribution to the data collection. This study was financially supported by the Dutch Ministry of Health, Welfare and Sport.

EuroQol Proceedings are available from: F. Th. de Charro PhD, EuroQol Group Business Manager, Centre for Health Policy and Law, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands.

PART II

Chapter 7

The Impact of Clinical Information on Patients' Preferences for Cancer Treatment:

An Exploratory Study based on Laryngeal Cancer

Paul F.M. Krabbe, Jos van der Donk, J. Dik F. Habbema,
Peter C. Levendag, Cees A. Meeuwis, Paul I.M. Schmitz
(submitted for publication)

Abstract

An experiment was conducted to study the effect of the level of detail of clinical information on preferences for treatment options.

The two predominant treatment modalities for T3 laryngeal cancer, radiation therapy (RT) and surgery (S), were valued by a sample of 48 participants. For each treatment modality, two different descriptions were used: a concise "Outcome" description and a more comprehensive "Course" description. These descriptions were assigned randomly to the participants. Valuations were elicited twice. In between, detailed additional clinical information about the treatment regimes was offered.

With the Course description, the two valuation of the treatments showed comparable means, which did not alter after offering additional information. Participants who valued Outcome descriptions generally preferred RT before additional information was offered, while afterwards values for RT and S were much closer to each other.

The level of detail of clinical information offered to patients may affect their treatment preference in a systematic way. This result has implications for the development and interpretation of preference measurements for clinical decision support.

Key words: clinical information, treatment preference, laryngeal cancer, QALEs, clinical decision making

Introduction

The measurement of patient preferences, which is a central issue in clinical decision support, is nowadays becoming even more important with the increasing emphasis on the patients' role in decision making. Provision of information and stimulating participation in choosing a management strategy are central in the enhanced patient role (Chapman *et al.*, 1995; Richards *et al.*, 1995). Whereas some patients and clinicians may give more significance to survival considerations, preferences of others may be more determined by expected "well being" or "quality-of-life" during and after a particular treatment regime. Individual patients' treatment preferences take shape during the patient-physician consultation and may depend on the type and the amount of information offered to the patient. A brief outline of the likely stable end-state after treatment could lead to an other treatment preference than a comprehensive description including explanation of possible side-effects and complications of the different treatment regimes.

The same dependency of preference and information could exist in the more formalized context of decision support, using methods from clinical decision analysis. In order to study this phenomenon, we took the context of treatment choice in T3 laryngeal cancer. This condition is predominantly treated by radiation therapy or by surgical removal of the larynx including the vocal cords. Controversy exists regarding the optimal treatment (van den Bogaert *et al.*, 1983; Levendag *et al.*, 1993). The best prognosis for survival is achieved through surgery, although life expectancy after radiation therapy is estimated to be only slightly lower than for surgery. Compared to surgery, treatment by radiation therapy has the advantage of preservation of speech and the preclusion of a stoma. The tradeoff between quantity and quality-of-life for this type of cancer can be studied using formal techniques based on decision analysis (McNeil *et al.*, 1981; Pauker & McNeil, 1981; Pauker & Kassirer, 1987; Elstein, 1989).

In an experimental study, valuations for the two treatments (radiation therapy and surgery) were obtained by offering subjects one of two specific descriptive formats with different types and amounts of clinical information. One concise format ("Outcome") was used for assessing values for the stable end-states of the treatments. These values were subsequently used in a clinical decision analysis. In order to compute Quality-Adjusted Life Expectancies (QALEs), they were combined with life expectancies and outcome probabilities. In the alternative, comprehensive, descriptive format ("Course"), treatments were assessed using a summary of the information that would typically be utilized in conventional (informal) decision making in the patient-physician encounter. Assessments for the descriptions were carried out two times. In between, detailed additional information was orally offered in order to study the effect of more extensive background knowledge.

By employing this approach we were able to address the following question: are there notable effects on the preferences for treatments related to the type and the amount of clinical information offered to respondents? And how are these preferences influenced by offering additional background information?

Subjects and Methods

Respondents

The total sample size was 48: 24 former laryngeal cancer patients and 24 lay volunteers. The former patients were free of the disease for at least three years. They were eligible if they had a Karnofsky score of more than 50 and had no psychiatric history. All former patients from the patient records of the Dr Daniel den Hoed Cancer Center in Rotterdam who satisfied these requirements received a letter from their clinician explaining the study. From this pool, former patients were randomly asked by telephone to cooperate in the study until 24 were entered. The lay volunteers were inhabitants of Rotterdam who reacted to an advertisement in a local newspaper in which healthy subjects of 50 years of age and older were invited to participate. From the 54 people who responded 24 were selected who matched the age and sex composition of the former patients.

Treatment outcomes and probabilities

For laryngeal cancer, especially for classification $T3N_0M_0$, there are essentially two treatment modalities: radiation therapy (RT) exclusively, and laryngectomy (i.e., surgery: S) with or without additional radiation therapy. There are obvious advantages and disadvantages to both treatments. RT has a lower life expectancy but with the opportunity to retain natural speech. S has a higher life expectancy, but with a stoma and loss of normal speech. Although saving natural speech, in practice RT involves more than just accepting a lower survival rate: in about half of the patients, the initial RT is followed by S after some time because of recurrence of the carcinoma.

The probabilities and years of survival used in this study stem from a Dutch multi-center study performed by the Dutch Cooperative Head and Neck Oncology Group (Levendag *et al.*, 1993). These results were entered in a decision tree, which contains the possible actions and outcomes (Figure 1).

Descriptions and elicitation method

We used two different descriptive formats to elicit scores for RT and S: "Outcome" descriptions and "Course" descriptions (see: Appendix 7.1). Outcome comprised a brief description (RT 35 words, S 79 words) with relevant qualitative information about the stable end-state after RT or S. The more elaborate Course format (RT 241 words, S 210 words) included the treatment program with prognostic length-of-life information, e.g., 5-year survival rates, and quality-of-life information, e.g., stable end-state, level of mutilation, physical impairment and side-effects.

The "Outcome" format is employed within the framework of clinical-decision analysis. The values – or utilities –, elicited for these outcome constitute one of the building blocks used for calculating the Quality-Adjusted Life Expectancy (QALE), a measure expressing the preferences for the treatments. (Weinstein *et al.*, 1980; Torrance & Feeny, 1989). An example

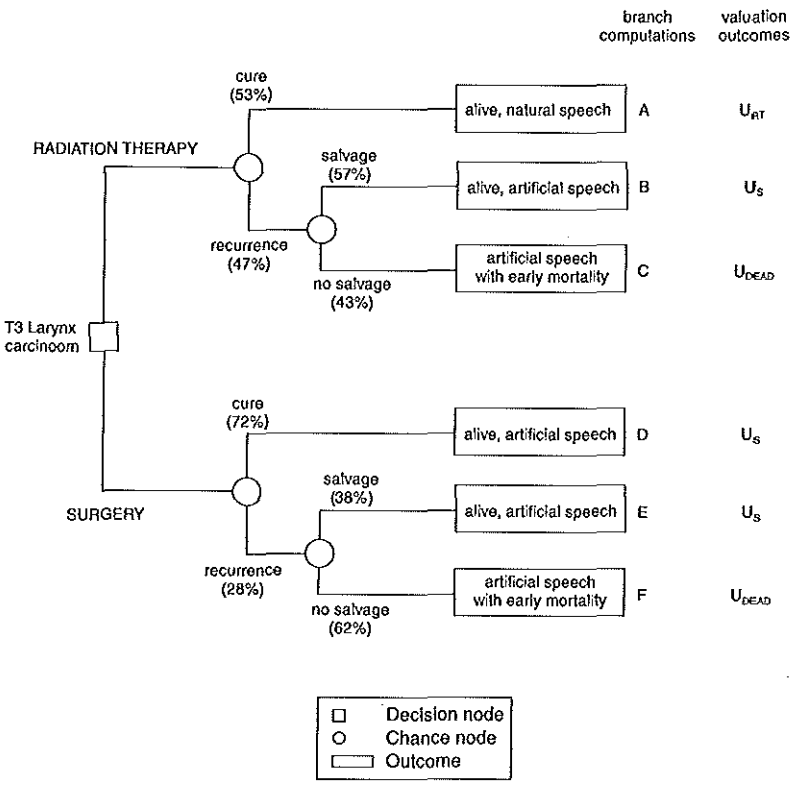


Figure 1
Decision tree for the choice (square) between the two treatments, i.e., radiation therapy and surgery for a 65-year-old man with $T_3N_0M_0$ laryngeal cancer. Probabilities of outcomes are displayed after each chance node. For detailed explanation see Appendix 7.2

of the calculation of QALEs for the laryngeal cancer problem is given in Appendix 7.2. The "Course" format, on the other hand, is based on a one-stage integral or holistic valuation of the three combined attributes of the QALE measure, i.e., outcomes, life expectancies and probabilities. The resulting QALEs express a preference for treatments.

Opinions on which method to use to elicit valuations differ (Froberg & Kane, 1989b; Nord, 1992). Three methods (time-tradeoff, standard-gamble, visual analogue scale) were applied in this study. The comparison of these three methods will not be discussed here. We will use the results of the time-tradeoff (TTO) because this method is feasible and was specifically developed for use in medical settings (Torrance, 1976).

In performing the TTO task, participants were presented with two alternatives and asked to select the preferred one. We specified the first alternative as a description (Outcome, Course) with a duration of 10 years. The competing alternative offered perfect health of

shorter duration. The duration (y) in perfect health was varied in steps of 0.5 years until the point of indifference between the two alternatives was reached. The valuation, calculated as $y/10$, was however discounted because time preference studies have revealed that people value later years as less valuable than earlier ones (Weinstein, 1993). To adjust for this, we assumed a discount rate of 3.5% for all valuations elicited by TTO and appropriately transformed the TTO valuations (Johannesson, 1995; Weinstein *et al.*, 1996).

Study design

After a first valuation interview intended to familiarize the participants with the assessment tasks, all 48 participants were interviewed twice during a second visit two weeks later. In between these two valuation interviews during the second visit, detailed additional clinical information (from now on referred to as "additional information") on T3 laryngeal cancer was offered. This additional information explained both treatment modalities and their possible outcomes in understandable terminology. The assessments before and after offering the additional information will be indicated respectively as the "pre-info" and the "post-info" interview. This part of the study was carried out by a trained interviewer (JvdD) who used visual and verbal techniques to offer the information. During this session, the participants were free to ask questions about all treatment-related aspects and other features pertaining to the topic. The 48 participants were randomly divided into two experimental groups of 24 participants, Group I and Group II, that assessed the Outcome and the Course descriptive formats respectively. The present study design was based on experiences gained during a pilot study (van der Donk *et al.*, 1995).

Analysis

Mean and standard deviations were computed for the Outcome-values and the Course-values. Subsequently, these valuations were statistically analyzed with MANOVA (repeated measurement design) for the factors "replication" (the two successive assessments: pre-info, post-info), "descriptive format" (Outcome, Course) and the factor "treatment" (RT, S). Correlation coefficients (intraclass correlation coefficients, ICC) were computed particularly to test the equivalence of the elicited valuations before and after the offering of additional information (Deyo *et al.*, 1991). For the Outcome descriptions, we computed QALEs through the multiplication of the Outcome-values with the other parameters (life expectancies, probabilities) in the decision tree.

Possible differences in preferences between the former patients and the lay volunteers will be discussed separately.

Results

Descriptive outcome

The ages of the former laryngeal cancer patients (M = 62.2, SD = 9.7) and the lay volunteers (M = 61.1, SD = 7.4) closely matched. Males were in the majority, as expected for this condition. In selecting the lay volunteers, this unequal distribution between males and females was reproduced.

Overall effect of additional information

Table 1 presents the means, difference scores and standard deviations for treatment valuation. Both groups valued treatment modality S lower than RT. Minor differences in average valuation before and after offering additional information can be observed in three of the four valuation tasks. In one valuation task, the RT modality based on the concise Outcome descriptions, changes proved to be statistically significant.

Figure 2 demonstrates the two main results graphically: 1) scores based on the Course descriptions yielded almost identical means for the two treatment options, 2) due to the offering of additional information, the preference "gap" between RT and S was diminished.

Multivariate analysis (MANOVA) revealed that of the three main factors (replication, descriptive format, treatment) only "treatment" was significant ($p < 0.01$), which means that

Table 1

Means and standard deviations (between parentheses) of the valuations elicited by the time-tradeoff method before and after offering detailed information (pre-info, post-info) by two different clinical decision approaches (Outcome, Course) for two treatment modalities (radiation therapy, surgery); valuations are on a scale from 0 (death) to 1 (perfect health)

Descriptive Format		Treatment Modalities	
		Radiation Therapy	Surgery
Group I (n = 24)			
Outcome description			
	pre-info score	0.85 (0.15)	0.66 (0.24)
	post-info score	0.79 (0.19)	0.68 (0.24)
	difference	-0.06* (0.14)	0.02 (0.09)
Group II (n = 24)			
Course description			
	pre-info score	0.70 (0.22)	0.65 (0.24)
	post-info score	0.69 (0.25)	0.66 (0.24)
	difference	-0.01 (0.10)	0.01 (0.08)

RT and S were valued systematically different by the participants. Of the 11 interaction terms two terms were significant. The most relevant interaction term in this context, "treatment × replication" ($p = 0.01$) reveals that the offering of clinical information affected systematically the valuation of RT in a different way than S. The interaction term "treatment × descriptive format" ($p = 0.02$) reflects that for the Outcome descriptions, RT was valued systematically higher compared to the Course descriptions, but S not.

Individual preferences

The valuations elicited by the Course descriptions produced almost similar values at the individual level on the pre-info and post-info interviews for both RT (Intraclass Correlation Coefficient, ICC = 0.90) and S (ICC = 0.95). For the Outcome descriptions the valuations of S were rather comparable on the two interviews (ICC = 0.93). However, much more dissimilar valuations were elicited between the pre-info and the post-info interviews by the Outcome description for RT (ICC = 0.66). This low coefficient can be a result of a major difference of the overall mean between the two valuation interviews, but can also indicate difference of valuations of the participants on the two interviews. Contrary to the conventional Pearson correlation coefficient, the ICC statistic is susceptible to group level shifts. Low coefficients therefore will result from a major difference of the overall mean between the two valuation interviews and/or a heterogeneous valuation of individuals between the two interviews. The individual preferences are depicted in Figure 3. On the pre-info interview the two descriptive formats both yielded a higher number of preferences for

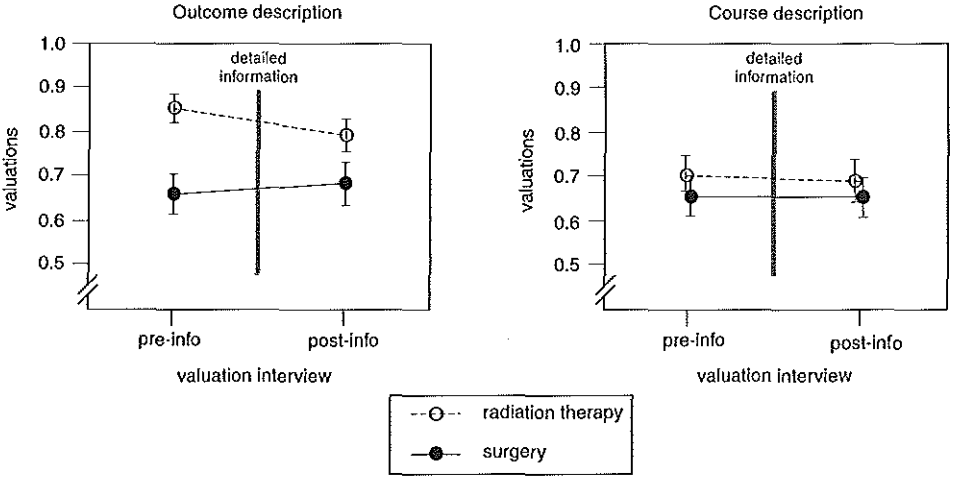


Figure 2
Mean valuations elicited by the time-tradeoff method before and after offering detailed information (pre-info, post-info) using two different descriptions (Outcome, Course) for the two treatment modalities radiation therapy and surgery (standard error of the mean are depicted)

RT. Five participants altered their preference due to the offering of additional information in the Outcome description group (their preferences all shifted from RT to S). As a result, we observed that a strong tendency to choose RT at the pre-info interview changed approximately into an equal number of preferences for both treatment modalities on the post-info interview.

For the Course descriptions, partially due to the additional "indifferent" preference option, more shifts results from offering additional information. In contrast to the Outcome descriptions, a predominant indifferent situation was observed for the Course descriptions at both interviews. However, the tendency to choose RT among the "non-indifferent" persons, was somewhat weakened after offering additional information.

Former patients versus lay volunteers

As far as descriptive statistics are concerned, the only distinction between the former patients and the lay volunteers was that the standard deviations of the difference in valuation between pre-info assessment and post-info was notably smaller for the former patients (not presented).

A MANOVA analysis in which the factor "group" (former patients, lay volunteers) was

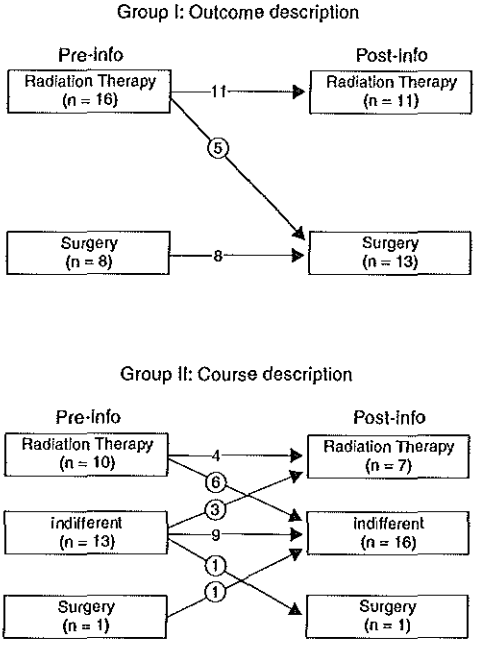


Figure 3
Individual treatment preferences based on the two descriptive formats (Outcome: via QALEs; Course: direct assessment) for the treatment modalities radiation therapy and surgery elicited at two valuation interviews (pre-info, post-info; shifts of preference are encircled)

also considered showed that this factor was significant ($p < 0.01$), because for the Course description, the former patients valued both RT and S higher (better) than the lay volunteers.

In comparison to the former patients, the lay volunteers expressed stronger preferences in favor of RT. This result applied for the preferences ($p = 0.08$) evoked by the Course descriptions as well as for the outcome-valued based QALEs.

Discussion

Patient involvement is especially important for choices in which patients are confronted with tradeoffs between survival and quality-of-life. One important aspect in patient couseling is the type and the amount of clinical information offered to the patient. We studied this aspect empirically for the two main optional treatment modalities of T3 laryngeal cancer, radiation therapy (RT) and surgery (S). Clear differences in response were found when varying the type and the amount of information.

The offering of detailed additional clinical information had no considerable influence on the valuations for participants who had valued descriptions which already comprised comprehensive information (Course descriptions). This was in contrast to the concise descriptions (Outcome descriptions), which comprised solely information about the stable end-states where there was such an effect. Due to the offering of additional information, participants in the Outcome group presumably became aware that if they chose RT, there was still a 47 percent chance of recurrence and the consequent surgical removal of the vocal cords. As a result, the initial values for RT diminished remarkably after offering additional clinical information. Though the (statistically significant) shift may look rather small, from 0.85 pre-info to 0.79 post-info (Table 1), understood as a loss compared to good-health, there is an inverse from 0.15 to 0.21, or a 40% increase, which is considerable. Likewise, side-effects and follow-up treatments of RT can be burdensome and patients may not be aware of them if only the stable end-states are described (Outcome). Offering additional information concerning these aspects may therefore result in a shift of preference. We hypothesized a priori that no significant effects would result from offering additional information for the group that assessed the Course description because the clinical information was already rather extensive. This hypothesis was confirmed by the study.

In order to compare individual preferences, the values of the concise Outcome description were incorporated into a decision analysis to estimate the preference measures of this analytical approach, i.e., Quality-Adjusted Life Expectancies (QALEs). Individual preferences based on these QALEs revealed that at the first interview RT was highly preferred. After the offering of additional information some participants changed their initial preference from RT to S: about half of the group preferred RT and the other half S. Because the Course valuation enables individuals to take an indifferent position, the offering of additional information effected no shift from one treatment to another, but only from a treatment preference to indifference vice versa. In fact, most of the individuals revealed themselves as "indifferent" to RT and S at both interviews.

We detected one significant effect between the former patients and the lay volunteers: former patients valued the Course descriptions systematically higher than the lay volunteers for both treatments. Adaptation processes by the former patients may be the cause for these differences.

Decision support approaches like clinical decision analysis or direct choice models use formal methods for quantifying preferences. In clinical decision analysis, valuation of outcomes is strictly separated from their duration and from the probabilities of their realization. Our study results concerning the influence of background information can evoke two opposite conclusions. The first one is that our findings prove that the strict separation should be maintained, because otherwise "double-counting" effects may occur. For example, when the information on the considerable likelihood of salvage surgery after recurrence is the main cause of the lower valuation of radiation therapy after the offering of additional information, we have with double-counting, because the same prognostic element is incorporated twice in the decision tree: as the implicit appraisal of this element in the assessment of the radiation therapy treatment description, and as the probability of recurrence in the decision tree. The opposite conclusion would be that the results prove the importance of background information in the valuation of health outcomes, and therefore should be provided to patients. If there would be double-counting when using decision analysis, this second conclusion implies that decision analysis should be adapted, and not the outcome valuation process.

Both reactions cannot be rejected outright, in our opinion. But there is a rather convincing "acceptability" reason for preferring the second reaction. Decision support will for the time being – and maybe always – be compared to "good clinical practice", also with regard to patient information. The acquirement of information over time is typical for clinical practice (discussion with physician, nurses and relatives; television programs; patient groups; information leaflets etc.). Consequently, patients will always have more knowledge about their optional treatments and their related consequences than only about treatment outcomes. Therefore, realistic clinical decision support will always have to assume a general background knowledge by the patient of the disease and its treatment, even if you offer solely outcome information.

When offering additional information before a second Outcome assessment task, a notable and statistically significant shift from preference of RT to S was observed. Because the additional information did not contain further information on the outcomes after treatment, the additional knowledge should be considered as "irrelevant" with respect to the Outcome valuation task. The significant shift in treatment valuation that occurred nevertheless must therefore be explained as a kind of "contamination" of the valuation task proper by background knowledge. Approximately 20% of the individuals made a shift from RT to S, and none the other way round.

Inevitably, performing evaluations based on Course descriptions demands a considerable cognitive task by participants, although it may, within a clinical counseling program, contribute to the active participation of patients in decision making. Because of the type of calculation involved, the QALE (Outcome) approach always results in a preference for either RT or S, whereas preferences elicited by Course type of descriptions may lead to indifferent situations. Indeed we found that most of the valuations based on the Course descriptions led

to indifference between RT and S. Offering additional information caused for 4 participants' changes, i.e., three from indifference to RT and one from indifference to S. But 7 participants also altered their pre-info preference for RT or S into indifference to these two treatments. Because of its high number of indifferent "preferences" it might be concluded that the Course approach poorly discriminates between the treatment options. An alternative conclusion is that for T3 laryngeal cancer, the indifference between RT and S is actually reflecting the apparent position of the participants, and that forcing preferences would only be misleading.

Some caveats considering the relation to clinical practice should be mentioned. First, it is important to remember that in clinical practice, patients generally receive treatment information predominantly verbally, with the possibility of feedback. In this study, the information was primarily offered in written format which may have introduced for some participants difficulties with understanding and cognitive processing. Second, in contrast to former patients and lay volunteers, actual patients facing a real-life decision about a life-threatening illness are under enormous emotional stress which may result in loss of recall or selective recall. In order to alleviate this stress, simple evaluation tasks such as outcome valuation in the context of formal decision analysis may be preferable, particularly as this approach does not confront patients with survival data.

In conclusion, the type and amount of information may affect the choice of treatment. This was the case in this study where as a result of offering additional information, a number of participants arrived at another preference or abandoned their initial indifference.

Acknowledgment

This study was supported by a grant from the Dutch Cancer Society (grant DDHK 91-05).

We express our gratitude to the Dutch Cooperative Head and Neck Oncology Group, Utrecht, the Netherlands, for enabling us to use some of their data on the outcome of T3 laryngeal cancer treated by either radiation therapy or surgery.

Appendix 7.1

Outcome descriptions

Radiation therapy

- You receive radiation therapy for a tumor on a vocal cord and because of this you have some problems (otherwise you are healthy).
- You have a hoarse voice, lack of appetite and a dry mouth.

Surgery (and eventually additional radiation therapy)

- You are operated on and receive additional RT for a tumor on a vocal cord and because of this you have some problems (otherwise you are healthy).

- You speak by means of a so-called speech button and you are reasonably understandable. Because during the operation a part of your windpipe was removed you have to breathe through an opening in the middle of your throat. You suffer from lack of taste and smell, extensive coughing and susceptibility to cold.

Course descriptions

Radiation therapy

- You receive radiation therapy for a tumor on a vocal cord. The treatment is given on a daily basis, is painless and has a duration of 7 weeks.
- To our knowledge after 5 years approximately 65% of people are still alive. As a result of the treatment, the following side effects are possible: skin on the neck might temporarily show symptoms of burning (like a heavy sunburn), general fatigue, and pain while eating and swallowing. In some cases, there may be a decrease in the production of sputum, which results in a permanently dry mouth.
- Natural speech is preserved, but the sound of the voice may change.
- If the tumor returns after RT (there is an approximately 50% chance that the tumor returns within 2 years), surgery will follow. During the operation, the larynx and the vocal cords will be removed and natural speech will be impossible.
- A speech button is implanted and soon you are able to speak coherently. By means of exercise you develop this speech and learn to speak in an alternative fashion, the so-called esophageal speech. After the operation, respiration through the mouth and/or the nose is no longer possible. It now occurs through a stoma in the neck.
- Permanent side effects of this treatment are loss of taste and smell, habitual coughing, and rapid susceptibility to cold. You have to attend the clinic on a regular basis for inspection, maintenance and exchange of the speech button.

Surgery (and eventually additional radiation therapy)

- You are operated on for a tumor on a vocal cord followed by additional radiation therapy. To our knowledge after 5 years approximately 70% of people are still alive.
- During the operation, the larynx and the vocal cords will be removed and natural speech will be impossible.
- A speech button is implanted and soon you are able to speak coherently. By means of exercise you develop this speech and learn to speak in an alternative fashion, the so-called esophageal speech. After the operation, respiration through the mouth and/or the nose is no longer possible. It now occurs through a stoma in the neck.
- Permanent side effects of this treatment are loss of taste and smell, habitual coughing, and rapid susceptibility to cold.
- About 1 month after the operation, additional radiation therapy is given. During this radiation therapy, x-rays are used to destroy remaining tumor cells. The treatment is given on a daily basis lasting five weeks.
- In some cases, there may be a decrease in the production of sputum, which results in a permanently dry mouth.

- There is a possibility of approximately 30% that the tumor will return within 2 years. You have to attend the clinic on a regular basis for inspection, maintenance, and exchange of the speech button.

Appendix 7.2

Quality-adjusted life expectancy

As shown in Figure 1 the treatment modalities for patients with T3 laryngeal cancer consist of radiation therapy (RT) and surgery (S). The expected effectiveness of the treatment modalities can be expressed as life expectancy (LE). In the specific case of T3 laryngeal cancer, the life expectancy for a successfully treated 65-year-old male patient is similar to the age-specific LE in the male Dutch population, which is 14.1 years. The life expectancy after unsuccessful treatment, leading to recurrence and early death is estimated to be 3 years (composed of two years with natural speech and one year with artificial speech).

In this study, the following rates for cure, recurrence and salvage were used as well years of survival (underlined), leading to the present computations (rates and years of survival stem from the Dutch multi-institutional study; see: Levendag *et al.*, 1993).

LE_{RT}	$(0.53 \times \underline{14.1})$	$= 7.47$	LE_{S}	$(0.72 \times \underline{14.1})$	$= 10.15$
	$(0.47 \times 0.57 \times \underline{14.1})$	$= 3.78$		$(0.28 \times 0.38 \times \underline{14.1})$	$= 1.50$
	$(0.47 \times 0.43 \times \underline{3.0})$	$= 0.61$		$(0.28 \times 0.62 \times \underline{3.0})$	$= 0.52$
	LE_{RT}	$=11.86$ years		LE_{S}	$= 12.17$ years

Based on life expectancies, surgery is the optimal treatment. Whereas health-related quality-of-life is ignored by the LE approach, it is embodied in the quality-adjusted life expectancy (QALE) measure.

The QALE makes adjustments by using individual valuations (utilities) of the possible health states. Its unit is the quality-adjusted life year (QALY). After treatment of T3 laryngeal cancer the patient is assumed to be in one of three possible health states: alive with artificial speech (utility U_S), alive with natural speech (utility U_{RT}) or death in the near future (utility U_{DEAD}). It is assumed that the utility for $U_{DEAD} = 0$ for both treatment modalities.

In the calculations below, we assume that the assessed utilities (printed in italics) for a particular patient are $U_{RT} = 0.93$ and $U_S = 0.75$, on a scale from 0 to 1 (perfect health). QALEs for the two treatments are:

$QALE_{RT}$		$QALE_S$	
$(0.53 \times \underline{14.1} \times 0.93)$	= 6.95 (A)	$(0.72 \times \underline{14.1} \times 0.75)$	= 7.61 (D)
$(0.47 \times 0.57 \times \underline{2.0} \times 0.93)$	= 0.50 (B)	$(0.28 \times 0.38 \times \underline{14.1} \times 0.75)$	= 1.13 (E)
$(0.47 \times 0.57 \times \underline{12.1} \times 0.75)$	= 2.43 (B)	$(0.28 \times 0.62 \times \underline{3.0} \times 0.75)$	= 0.39 (F)
$(0.47 \times 0.43 \times \underline{2.0} \times 0.93)$	= 0.38 (C)		
$(0.47 \times 0.43 \times \underline{1.0} \times 0.75)$	= 0.15 (C)		
$\Rightarrow$ $8.41 \times 0.93 = 7.82$		$\Rightarrow$ $12.17 \times 0.75 = 9.13$	
$3.45 \times 0.75 = 2.59$			
$QALE_{RT}$ = 10.41 QALYs		$QALE_S$ = 9.13 QALYs	

In this example, RT is thus the preferred treatment on the basis of the QALE, whereas S was preferred when using the LE.

Chapter 8

Analytical versus Holistic Approaches to the Appraisal of Treatments

Paul F.M. Krabbe, J. Dik F. Habbema, Peter C. Levendag,
Cees A. Meeuwis, Jos van der Donk, Gouke J. Bonsel, Paul I.M. Schmitz
(submitted for publication)

Abstract

In order to compare two different approaches (analytical, holistic) to determinate treatment preferences, we examined the choice between radiation therapy and surgery in T3 laryngeal cancer, using two comparable groups of respondents.

In the analytical approach, individual values for each separate treatment outcome were elicited from one group of respondents. Subsequently, these values were integrated into a conventional decision tree analysis, that is, weighted by probability of occurrence and life expectancies. In the holistic approach, treatment preferences were elicited after offering respondents all relevant medical information pertinent to the effects of the complete course of each treatment. No further steps were necessary for this second approach. Respondents were 24 former laryngeal patients and a matched group of 24 lay volunteers, randomized over the two experimental groups. Preferences for all assessments were elicited by three different elicitation methods; a final unweighted forced-choice was added.

Responses at the group level were almost similar for both approaches. At the individual level, the preferences for one of the two treatments, however, varied substantially between the two approaches for all elicitation methods. The correspondence between individual preferences in the analytical approach and their forced-choice for one of the treatments was high.

We conclude that the validity of the analytical approach in the appraisal of treatments was reasonably sustained.

Key words: preferences, clinical decision making, analytical approach, holistic approach, decision tree analysis, descriptive formats

Introduction

The choice of treatment for an individual patient frequently involves tradeoffs between competing health aspects as well as other personal considerations. There have been increasing demands for more patient involvement in decision making and a growing recognition by health care providers that such participation is desirable. Within this context, the development of clinical decision analysis, a scientific approach that assesses the preferences of individual patients for treatment options, has been viewed as attractive. Despite the availability of this framework, the incorporation of patient preferences is still rare. The majority of studies use volunteers and former rather than current patients (McNeil *et al.*, 1981; O'Connor *et al.*, 1987; Maas & Stalpers, 1992; Stiggelbout *et al.*, 1994). Moreover, the framework has been criticized for several reasons. First, although treatment outcome is considered to be the most important attribute, it is not the only determinant of preference. Discomfort due to diagnostic procedure, complications of drug therapy (Kassirer, 1983) or necessary investment of physical (rehabilitation efforts), mental or financial resources (Hilden *et al.*, 1992) are important "process" aspects (Donaldson & Shackley, 1997). Second, individual decision analysis is based on a set of axioms and assumptions which have been theoretically challenged (Schoemaker, 1982; Broome, 1993; Richardson, 1994; Cohen, 1996; Nease, 1997). Finally, the empirical validity of the approach has been questioned, i.e., is this approach capable of providing a valid and consistent representation of individual preferences for medical outcomes?

The decision problem we studied consisted of the choice between the two treatment modalities for laryngeal cancer, classification $T_3N_0M_0$, i.e., radiation therapy (RT) exclusively, or laryngectomy (i.e., surgery: S) with or without additional radiation therapy. The two treatments and their outcomes can be represented as a decision tree (see: Figure 1). The life expectancy for a successfully treated 65-year-old male patient is 14.1 years, similar to the age-specific life expectancy of the male Dutch population. After unsuccessful treatment, life expectancy is estimated to be three years (two years until recurrence and one year between recurrence and death). If we include the associated probabilities for both options, RT on balance has a slightly lower life expectancy (11.9 years) but provides an opportunity to retain natural speech; S offers a higher life expectancy (12.2 years), but causes the loss of normal speech and requires a stoma.

We compared two competing approaches of representation and valuation of clinical decision problems in order to investigate opportunities of obtaining valid preferences in patients with laryngeal cancer. The *analytical* approach consisted of definitions of treatment outcomes, that is, two sets of stable end-states. These outcomes were subsequently valued by three elicitation methods (standard-gamble, time-tradeoff, visual analogue scale). The resulting values were combined with survival and probability data to calculate a standard Quality-Adjusted Life Expectancy (QALE) for both treatment options. The analytical approach uses clinical analysis to attain a treatment choice. The *holistic* approach uses a comprehensive clinical treatment and course description for each of the two treatment options. The latter approach provides preferences which, without the demand of additional

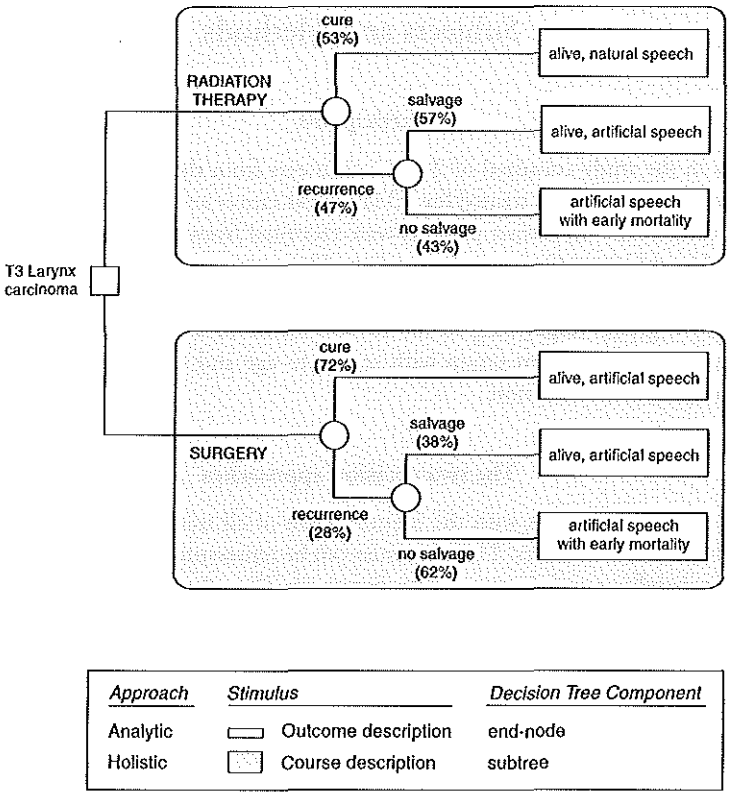


Figure 1

Decision tree for the choice between radiation therapy and surgery for a 65-year-old man with $T3N_0M_0$ laryngeal cancer; the information represented by the tree that forms part of the descriptions for the analytical and the holistic approach are both graphically highlighted

computations, are supposed to represent an analogous process as the analytical calculation of QALEs. Again the same three elicitation methods were used.

Our main purpose was to compare the two different approaches in the assessment of patient treatment preferences.

Subjects and Methods

Participants

Participants in the study consisted of 24 lay volunteers from the Rotterdam population and 24 participants who had been treated for T3 laryngeal cancer and who were free of the disease for at least three years. All former patients were recruited from the patient records of the Dr Daniel den Hoed Cancer Center in Rotterdam. They were randomly asked by telephone to co-operate in this study until 24 former patients were entered. The lay volunteers were selected from a group responding to an advertisement in a local newspaper. They were matched to the former patient group with regard to age and sex, as laryngeal cancer has a specific demographic distribution.

The two decision approaches

The *analytical approach* starts with the identification of all transitory and permanent states following the treatment choices. This outcomes can be represented in a decision three with the associated probabilities and durations (see: Figure 1). The relative worth of each permanent state, referred to as *utility* but also referred to as value or weight, has conventionally to be supplied by the patient him/herself. After treatment of T3 laryngeal cancer, the patient is assumed to be in one of three possible permanent/stable health states (see: Figure 1) i.e., alive with natural speech (utility U_{RT}), alive with artificial speech (utility U_S) or death in the near future (utility U_{DEAD}). It was assumed that the utility for $U_{DEAD} = 0$ for both treatment modalities. The two utilities, one for radiation therapy (RT) and one for surgery (S), were provided by the respondents responding on the valuation task (see below).

The final step was the calculation of the Quality-Adjusted Life Expectancy (QALE) for each treatment strategy. The expected time periods of different outcomes and their corresponding probabilities were for each treatment option separately multiplied by the respective individual utilities (U_{RT}, U_S). Results expressed in QALEs are indicated as *QALY-based preferences.*

The *holistic approach* is based on descriptions of the complete course of treatment, comprising the three attributes of the QALE measure (i.e., clinical state, durations and the respective probabilities). Its values are indicated as *holistic preferences.*

Descriptive formats

Each approach used a specific descriptive format or style. The analytical approach applied the conventional "shorthand" style, while the holistic approach used a narrative style (Torrance, 1976; Bonsel *et al.*, 1992) (see: Appendix 8.1). The brief description (RT 35 words, S 79 words) of the analytical approach comprised relevant qualitative facts about the stable health state after RT or S. The narrative course description (RT 241 words, S 210 words) contained the treatment process complete with relevant medical elements and prognostic indications. More specifically, this description consisted of two kinds of information: quantitative

information (e.g., tumor recurrence rates, 5-year survival rates) and qualitative information (e.g., type of mutilation, temporary and permanent side-effects, symptoms and physical impairment). Hence, only global medical information related to the two treatments was offered and there was a deliberate decision not to offer any information about possible impacts on the psycho-social domain. Both type of descriptions were presented on paper.

The probabilities and years of survival in this study were derived from a Dutch multi-center study (Levendag *et al.*, 1993) whereas the descriptions of health status and clinical states were based on previous publications on this subject and information from the clinicians.

Elicitation methods

Three elicitation methods were used, the same for both approaches: standard-gamble (SG), time-tradeoff (TTO) and the visual analogue scale (VAS). In absence of a gold standard (Nord, 1992; Krabbe *et al.*, 1997) we used the three most frequently applied methods. The SG and TTO methods are so-called tradeoff methods and were developed in the field of health economics (Torrance *et al.*, 1972; Weinstein *et al.*, 1980; Torrance, 1986). The VAS is a traditional method which is used in many fields for valuing subjective phenomena.

SG uses a probability concept and is offered as an iterative paired comparison (Torrance *et al.*, 1972; Gafni, 1994). Participants are presented with two alternatives and asked to select the preferred one. We specified the first alternative either as an analytical (shorthand) or holistic (narrative) description of treatment outcome, while the competing alternative was specified as a gamble with mutually exclusive probabilities for two reference outcomes, i.e., a good health state and a worse health state (the same text in both approaches). The reference outcomes in our study were perfect health with probability (p), or death with probability ($1-p$). By varying the p-level of the uncertainty outcome in steps of 5%, the point of indifference between the two alternatives was determined (valuation = p). The first bid for the reference outcomes was the same for all treatment outcomes to be valued, viz., a 50% chance of being in perfect health and 50% chance of being dead.

The TTO method uses a time concept and is also based on an iterative procedure with two alternatives. We specified the first alternative as either an analytical or holistic description of either RT or S, with a duration of 10 years.* The competing alternative offered a perfect health state of shorter duration. The point of indifference between the two alternatives was reached by varying the duration (y) in perfect health in steps of 0.5 years (valuation = $y/10$). In our application the first bid of the perfect health state was stated as 5 years. Due to the fact that people generally value later years as less valuable than earlier years, valuations elicited by the TTO task are biased downwards, especially the intermediate valued descriptions (Johannesson *et al.*, 1994). To correct for this time preference effect (Lipscomb, 1989; Krabbe & Bonsel, 1998) we therefore assumed in the analysis a discount rate of 3.5% for all

* A time period equal to the empirical life expectancy in this case approximately 14 years, could have been used. In our opinion the advantage of this slightly better fit to the "real-life" data does not outweigh the cognitive advantage to tradeoff the two different treatments in terms of a 10-point scale.

valuations elicited by TTO (Redelmeier & Heller, 1993; Weinstein, 1993; Weinstein *et al.*, 1996).

The third method, VAS, requires the participants to value a particular health state by placing a mark on a 100 mm. line with two descriptive anchors at both end-points (Froberg & Kane, 1989b; Wewers & Lowe, 1990). In this study, "best imaginable health state" and "worst imaginable health state" were used as anchors (valuation = distance in mm. from "worst imaginable health state" to mark / 100). The choice of a different lower anchor-point for the VAS does not limit the comparison of the two approaches in assessing patient treatment preferences.

In the study design (see below) the order of the methods used for the elicitation was completely randomized for the participants.

Study design

All 48 participants were interviewed during three visits with two two-week intervals in between. Two groups were constructed by randomly dividing the former patients and the lay volunteers into two groups. Group I (n = 24) valued analytical descriptions while group II (n = 24) valued holistic descriptions. At the first visit, one complete valuation interview was held to familiarize the participants with the valuation interview. At the second visit, two identical valuation interviews were performed. In between information was offered as an experiment (data not shown here). During this second visit the participants were thoroughly informed and could ask additional questions about treatment-related aspects and other relevant features. The fourth and fifth interview held on the third visit were identical to the previous valuation interviews except that we presented the opposite descriptive format (cross-over) on the fifth interview for each of the two groups (see: Figure 2). After restating the most salient aspects of the two treatments and their differences, a direct forced-choice between RT and S was requested at the end of the fifth interview.

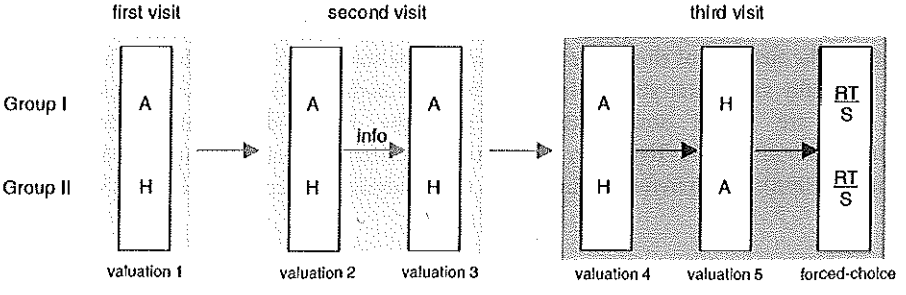


Figure 2
Schematic representation of the study design (A = analytical approach, H = holistic approach)

The results of these last two valuation interviews and the forced-choice of the third visit are discussed in this paper. The participants should be regarded as well informed, to the extent that their information level might correspond to the real-life situation of patients who are actually being asked to decide which treatment option they prefer.

Analysis

For the values (range 0.0 – 1.0) directly obtained after the assessment at the last two interviews (prior to the QALE computations for the analytical approach) we carried out the following analyses. Means and standard deviations were computed for the six subgroups separately (two approaches, three elicitation methods). Individual difference scores between RT and S were computed, averaged for all six subgroup comparisons and statistically tested (paired t-test). To test concordance between the values of the two approaches, we computed Intraclass Correlation Coefficients for RT and S (six subgroups). ICCs takes into account the variability due to two systematic sources of bias, i.e., a level effect between two measures and a linear transformation similar to Pearson correlation coefficients. The assumption that there is no relative difference between the values of the analytical approach and the values of the holistic approach was tested by non-parametric tests (Wilcoxon).

For the analytical approach, preferences (range 0.0 – 12.17) were expressed in QALEs based on conventional decision analysis. Analogous to the correction for the time preference produced by the TTO method, a similar discount rate of 3.5% was used for all durations in QALE calculations. Comparison of the QALE of RT and S defined preference for either option. Correcting by a discount rate was regarded inappropriate for the holistic approach, since all the factors influencing the individual evaluation, including the assumed time preference, were jointly accounted for by cognitive appraisal of the information.

Results

Participants

Three former patients refused to participate in the study. Two of them revealed reluctance to be confronted with the disease again, the other person had trouble with understanding the task in general due to his age. The ages of the former laryngeal cancer patients (M = 62.2, SD = 9.7) and the lay volunteers (M = 61.1, SD = 7.4) closely matched. Nine former patients had been treated by radiation therapy (RT), and the other 15 by surgery (S; sometimes plus radiation therapy). The distribution of the tumor stages of the former patients were as follows: ten stage II, ten stage III and four stage IV. Males were in the majority, as expected for this condition. In selecting the lay volunteers, this unequal distribution between males and females was reproduced.

Average values for RT and S in the six subgroups

Table 1 shows the means and standard deviations of all valuations in six subgroups. Radiation therapy (RT) was valued higher than surgery (S). The effect of change in the descriptive format on the valuations was much larger for RT than for S. For both RT and S, values elicited by the standard-gamble (SG) method proved to be systematically higher than those of the time-tradeoff (TTO) method, whereas valuations for the visual analogue scale (VAS) were lowest. Differences between RT and S were substantially smaller for the holistic approach than for the analytical approach. Five of the six tests for this latter approach proved to be significant ($p < 0.01$). The standard deviations were approximately the same under all conditions.

Lay volunteers only differed from the former patients in their slightly lower assessment of the holistic approach (not shown in table).

Table 1

Mean (standard deviations) of the valuation of stable end-states (analytical approach) and the overall valuation of the therapies (holistic approach). Group I first used the analytical approach and then the holistic approach; in group II the order was the reverse. Valuations were elicited on a 0 to 1 scale by three methods (standard-gamble, time-tradeoff, visual analogue scale) for two treatment modalities (radiation therapy, RT; surgery, S)

Elicitation Methods and Treatment Modalities	Descriptive Format			
	Group I (n = 24)		Group II (n = 24)	
	Analytic (first stimulus)	Holistic (second stimulus)	Analytic (second stimulus)	Holistic (first stimulus)
Standard-Gamble				
Radiation Therapy	0.87 (0.15)	0.74 (0.21)	0.81 (0.26)	0.72 (0.22)
Surgery	0.76 (0.20)	0.75 (0.21)	0.72 (0.25)	0.68 (0.22)
difference (RT - S)	*0.11 (0.19)*	*-0.01 (0.13)*	*0.09 (0.13)*	*0.04 (0.12)*
Time-Tradeoff				
Radiation Therapy	0.86 (0.12)	0.70 (0.21)	0.77 (0.26)	0.68 (0.22)
Surgery	0.70 (0.20)	0.67 (0.21)	0.71 (0.23)	0.68 (0.22)
difference (RT - S)	*0.16 (0.19)*	*0.03 (0.08)*	*0.06 (0.11)*	*0.00 (0.07)*
Visual Analogue Scale				
Radiation Therapy	0.76 (0.17)	0.53 (0.21)	0.70 (0.24)	0.61 (0.24)
Surgery	0.53 (0.25)	0.52 (0.22)	0.64 (0.25)	0.61 (0.25)
difference (RT - S)	*0.23 (0.30)*	*0.01 (0.16)*	*0.06 (0.22)*	*0.00 (0.12)*

The concordance between the individual values (see: Table 2) of the analytical approach and those of the holistic approach were moderately high, with an exception for RT as assessed by group I. The rank-order Wilcoxon test disclosed significant results for all the RT assessments.

Preferences at the group level

In Figure 3, the mean Quality-Adjusted Life Expectancies (QALEs) for the analytical approach and the mean values of the holistic approach are depicted on comparable scales. The order of valuation of the two descriptive formats clearly affects the results. Group I, who first performed the interview using the analytical approach, revealed QALEs with clear differences between the two treatment modalities, whereas such differences were much smaller for group II.

Individual preferences

At the individual level, the number of preferences in favor of one of the two treatments differed for the two approaches. Radiation therapy was considered more preferable by both groups when using the analytical approach (see: Table 3). If a participant showed no specific preference for either RT and S, based on the values of the analytical approach, S becomes the preferred option after the QALE computation. This is fully attributable to the longer survival after S for laryngeal cancer. It should be noticed that for a more meaningful comparison

Table 2
Concordance (intraclass correlation coefficient) between the analytical and the holistic valuation. Group I first used the analytical approach and then the holistic approach; in group II the order was the reverse (radiation therapy, surgery)

Elicitation Methods and Treatment Modalities	Group I (n = 24)	Group II (n = 24)
Standard-Gamble		
Radiation Therapy	0.42**	0.78**
Surgery	0.84	0.89
Time-Tradeoff		
Radiation Therapy	0.19**	0.81**
Surgery	0.86	0.86
Visual Analogue Scale		
Radiation Therapy	0.15**	0.72*
Surgery	0.73	0.74

* Significant differences for the two approaches: $p < 0.05$ (Wilcoxon statistic, non-parametric rank test for matched-pairs.
** Idem: $p < 0.01$.

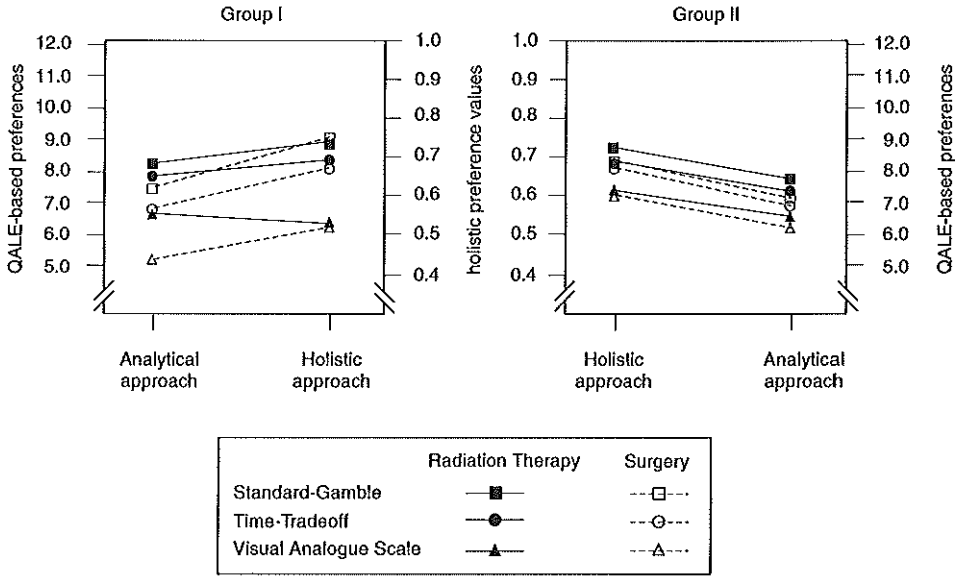


Figure 3
Mean preferences for the analytical and the holistic approach as obtained by three elicitation methods (standard-gamble, time-tradeoff, visual analog scale) for two treatment modalities (radiation therapy, surgery) for the two study groups (group I: first analytic then holistic, group II: first holistic then analytic). Computation of QALEs by decision tree analysis using a 3.5% discount rate for expected survival duration

between the two approaches we divided the indifferent scores of the holistic approach on a 50-50 basis to the two treatment modalities. By performing such an adjustment, we can see that the distributions for RT and S are almost the same for both approaches.

If asked to choose directly between the two therapies, 17 respondents in group I preferred RT (7 preferred S) and in group II 19 respondents preferred RT, see Table 3. For the forced-choices only a close resemblance could be observed with the number of preferences for the TTO and VAS methods for group I (resp. 17 and 18 for RT).

At an individual level, it appeared that there was a good correspondence between forced-choice and the analytical approach for group I: the SG, TTO and VAS resulted in respectively 17, 18 and 19 identical preferences. For the other three comparisons using the forced-choices (e.g., group I: holistic, group II: analytic + holistic) the correspondence was about chance (Md. = 12, range = 10 – 14).

Table 3

Individual preferences obtained by the valuation of stable end-states (analytical approach; QALEs after computation by decision tree analysis using a 3.5% discount rate) and the overall valuation of the therapies (holistic approach). Group I first used the analytical approach and then the holistic approach; group II first used the holistic approach and then the analytical one. Valuations were elicited by three methods (standard-gamble, time-tradeoff, visual analogue scale) for the two treatment modalities (radiation therapy, surgery); at last a forced-choice was requested

Elicitation Methods and Treatment Modalities	Group I (n = 24)		Group II (n = 24)	
	Analytic	Holistic	Analytic	Holistic
Standard-Gamble				
Radiation Therapy	12	5	13	9
Surgery	12	7	11	3
Indifferent	-	12	-	12
Time-Tradeoff				
Radiation Therapy	17	6	11	5
Surgery	7	3	13	5
Indifferent	-	15	-	14
Visual Analogue Scale				
Radiation Therapy	18	12	14	9
Surgery	6	11	10	12
Indifferent	-	1	-	3
Forced-Choice				
Radiation Therapy	17		19	
Surgery	7		5	

"-" = Indifferent by definition.

Discussion

Our study showed similarities between the holistic and the analytical approach at group level. The ranking of the two treatment modalities was on average the same for both approaches and concordance between the three elicitation methods could be observed. In the analytical approach, there was a strong treatment preference for RT, in particular if the analytical approach was performed first, whereas in the holistic approach this preference was less consistent. At the individual level, the correspondence between the preference for the analytical approach and the forced-choice was definitely higher than that between the holistic approach and the forced-choice.

In the analytical approach, only key aspects about the end-state are outlined, whereas the holistic approach confronts patients with unstructured multiple attributes. The latter

approach is characterized by its offering of process information, i.e., it presents a full picture of the treatment over time.

Several mechanisms may be responsible for the differences found between the two approaches, and depending on the *intended* conception of individual decision making, the holistic or the analytical approach may be preferred. We expect that comprehensiveness and presentation have influenced the results (Torrance, 1982; Llewellyn-Thomas *et al.*, 1984; O'Connor *et al.*, 1987; O'Connor, 1989). Remarkably, the effect of presentation order proved to be asymmetric. If we started with detailed (holistic) descriptions, the preferences based on the concise descriptions (analytical) were influenced more than when they were presented in the reverse order. Memory effects in regard to the holistic presentation may be responsible for this carry-over effect. The detailed information of the course of events may also focus the respondents more precisely on the consequences of both treatments. This profound introspection may decrease prior differences between the two treatment modalities, which may explain the smaller differences for the holistic preferences. Additional research should clarify the separate role of process information. If process effects exists, the validity of inclusion of process information will depend on the *intended* concept. From a narrow utilitarian perspective the incorporation of process information is seen as invalid. Within a broader view, however, the inclusion of process information seems justified.

Additional to conceptual considerations, the following arguments may guide the choice between the analytical or the holistic approach.

- An objection to the holistic approach is that it requires patients to remember and utilize too much data (cognitive overload) about the treatment outcomes in order to arrive at a consistent, well-considered valuation. As a consequence, patients may unintentionally focus on some direct and short-term effects instead of on the more abstract notions related to the duration of life and/or the probabilities associated with the health states. In addition, patients may overrate certain process elements, underrating the long-term impact and consequences of the possible permanent health outcomes.
- Valuations based on a holistic approach cannot be incorporated into the conventional concept of Quality-Adjusted Life Expectancies (QALEs), because the time-element and the associated probabilities of the transition from one state to another state are already included. Therefore, identical assessments based on the holistic approach cannot be used in studies that are alike for the treatment outcomes but with different treatment courses, whereas such a strategy is allowed for in the assessment of end-states in the analytical approach.
- Many aspects of the holistic approach may be similar to the decision making process of clinicians, although clinicians may be more inclined to emphasize the likely long-term outcomes. This can be understood from the broad experience of clinicians at each possible stage of a disease.

So far, the analytical and the holistic approaches apparently represent two extremes within the possible alternatives regarding the appraisal of treatments. Intermediate approaches have been suggested (Mehrez & Gafni, 1989; Hadorn *et al.*, 1992).

Despite the increased interest in incorporating individual patient's preferences into clinical decision making, many aspects affecting people's expression of their preference remain unclear

(Eraker & Polister, 1982; Torrance, 1987). Preferences may be sensitive to factors unrelated to the analytical/holistic approach, such as personal experience with particular treatments (Tversky & Kahneman, 1981) or specific diseases. In this study, however, the assessments of the former patients and the lay volunteers were highly comparable, except for a minor level effect. We therefore think that when panels are informed well, their background is irrelevant.

At the onset of the study we were aware of the possibility of elicitation method effects. Many investigators have reported that the methods that are used to elicit such preferences produce essentially different valuations (Torrance, 1976; Read *et al.*, 1984; Kaplan *et al.*, 1993). Alternatively, others have stated (Froberg & Kane, 1989b; Krabbe *et al.*, 1997) that the three elicitation methods give rise to comparable ordering of alternatives, albeit at different levels. The latter assertion seems to have been confirmed in this study.

Both the analytical and the holistic approach yielded reasonably consistent results at the group level, although radiation therapy was more clearly preferred in the analytical approach. Although the analytical approach requires further refinement and development, for example the assessment of intermediate states, only this approach showed high correspondence with the forced-choices for one of the treatments. Based on these results we therefore conclude that the validity of the analytical approach was reasonably sustained, whereas the holistic approach with its greater emphasis on patient participation in clinical decision making is worth studying, preferably using patient populations. Presumably, the participation of patients can be enlarged by the use of a holistic-like approach, conducted within the context of a well developed counseling procedure.

Acknowledgment

This study was supported by a grant from the Dutch Cancer Society (grant DDHK 91-05) and the Johan Vermeij Foundation.

We express our deep gratitude to the Dutch Cooperative Head and Neck Oncology Group, Utrecht, the Netherlands, for enabling us to share their experience and to allow the use of their data on the outcome of T3 laryngeal cancer treated by either radiation therapy or surgery. Furthermore, we would also like to thank SG Pauker for giving us a valuable suggestion at the 18th Annual Meeting of the Society of Medical Decision Making (Phoenix, Arizona).

Appendix 8.1

See Appendix 7.2 of Chapter 7.

Chapter 9

The Stability of Outcome Preferences:
A Study in Laryngeal Cancer

Paul F.M. Krabbe, J. Dik F. Habbema, Peter C. Levendag, Cees A. Meeuwis, Paul I.M. Schmitz
(submitted for publication)

Abstract
In an experimental setting we investigated the stability of preferences, by valuing two treatment outcomes for laryngeal cancer four times.

Assessments were performed by twelve former laryngeal cancer patients and by twelve healthy volunteers. Three valuation methods were applied: standard-gamble, time-tradeoff and the visual analogue scale. Concordance between the preferences was measured and significant effects of the experimental factors were examined.

Considerable differences between the consecutive valuations were found. Preferences obtained at the first session proved to be very unstable. Conversely, on the last assessment the preferences of the former patients and the healthy volunteers became almost comparable for all three methods. The standard-gamble method proved to be the most stable method for individual assessments.

Replication of the rather complex task of valuing treatment outcomes and rehearsal of the offered information seems necessary in order to arrive at stable preferences.

Key words: stability, preferences, utilities, medical decision making, valuation methods, laryngeal cancer

Introduction

Measures expressing the value assigned to particular health states are used in several areas in health care, ranging from the assessment of an individual patient's preference for treatments in clinical decision-analysis to planning and monitoring of health care programs. Such values reflect the level of distress, dislike, worth or desirability that subjects associate with a particular health state.

Clinical decision analysis, a formal method of optimizing individual choices is of growing importance because of the increasing involvement of patients in complex treatment decisions. The method presupposes valid individual assessment of the worth of treatment outcomes. These assessments usually take the form of valuation of descriptions of the stable end-states after treatment. Subsequently, the elicited preferences (values or utilities) are multiplied with other parameters (life expectancies and outcome probabilities) to compute Quality-Adjusted Life Expectancies (QALEs). In principle, the treatment with the highest QALE is expected to be preferred (Weinstein *et al.*, 1980; Pauker & Kassirer, 1987).

An important though neglected aspect in the field of health-state assessment is the stability or reproducibility of individual preferences elicited for treatment outcomes. A few studies have focused on the stability of individual preference assessment. Llewellyn-Thomas *et al.* (1984b) found that patients' utilities are reliable and stable, at least in the short term. But another study however produced opposite results (Christensen-Szalanski, 1984).

Different valuation methods exist for the elicitation of preferences (Torrance, 1976; Nord, 1992). The standard-gamble (SG) has the most direct links to Expected Utility Theory (EUT). There is an ongoing discussion about the merits of SG in relation to alternative methods (Torrance, 1986; Froberg & Kane, 1989b; Krabbe *et al.*, 1997). For example, the time-tradeoff (TTO) method has the advantage of being easier to understand. However, the TTO, like the other alternatives to SG, is not directly embedded in EUT. Nevertheless, in a number of empirical studies the TTO method yields comparable results, though generally slightly lower than SG (Torrance, 1976; Read *et al.*, 1984; Stiggelbout *et al.*, 1994; Bosch & Hunink, 1996; Krabbe *et al.*, 1996). Moreover, many methodological aspects concerning SG, TTO and other valuation methods are in debate (Mehrez & Gafni, 1991; Nord, 1992; O'Brien & Viramontes, 1994; Wakker, 1996).

The objective of this paper is to study the stability of preferences. In our study the preferences of treatments for stage T3 laryngeal cancer were examined. There were two optional treatments, radiation therapy and surgery (removal of the larynx including the vocal cords). Individual choice between radiation therapy and surgery primarily involve a tradeoff between survival and "quality of life". Surgery causes the loss of normal speech but results in higher cure rates than radiation therapy (van den Bogaert *et al.*, 1983; Levendag *et al.*, 1993).

Both treatments were assessed four times by three different valuation methods: SG, TTO and the visual analogue scale (VAS). It was a priori hypothesized, based on results from other studies, that SG and TTO would yield almost similar utilities, while those of VAS were expected to be generally lower. Our main aim however was to address the question whether preferences are sufficiently stable to justify the standard practice which is based on applying preferences which are elicited on a single occasion.

Material and Methods

Participants

Two groups of participants were enrolled in the study. One group consisted of 12 former patients who had been treated for T3-laryngeal cancer and were free of this disease for at least three years. They were considered to be eligible if they had a Karnofsky score of more than 50 and had no psychiatric history. All former patients from the patient records of the Dr Daniel den Hoed Cancer Center in Rotterdam who satisfied these requirements received a letter from their clinician explaining the study. From this pool, patients were randomly asked by telephone to cooperate, until the intended number of 12 former patients had been enrolled.

The other group consisted of 12 lay volunteers from the Rotterdam population who reacted to an advertisement in a local newspaper in which healthy subjects of 50 years of age and older were invited to participate in the study. Twelve respondents who reasonably matched the age and sex composition of the group of former patients were selected.

Treatment descriptions

Treatment descriptions for radiation therapy (RT) and surgery (S) were carefully composed, based on experiences from a preliminary study (van der Donk *et al.*, 1995), detailed discussion with the physicians and study of relevant publications. The short "naturalistic" descriptions (RT 35 words, S 79 words) contained only information about the probable stable end-state after RT or S. Deliberately, no prognosis statements, nor "process" aspects (Donaldson & Shackley, 1997), nor any reference to prognostic differences between RT and S were included.

Radiation therapy

- You receive radiation therapy for a tumor on a vocal cord and because of this you have some problems (otherwise you are healthy).
- You have a hoarse voice, lack of appetite and a dry mouth.

Surgery

- You are operated on and receive additional RT for a tumor on a vocal cord and because of this you have some problems (otherwise you are healthy).
- You speak by means of a so-called speech button and you are reasonably understandable. Because during the operation a part of your windpipe was removed you have to breathe through an opening in the middle of your throat. You suffer from lack of taste and smell, extensive coughing and susceptibility to cold.

Valuation methods

Three frequently used methods for eliciting preferences were used in this study: standard-gamble (SG), time-tradeoff (TTO) and the visual analog scale (VAS). The SG and TTO methods are so-called tradeoff methods and were originally developed in the field of

economics. The VAS scale is a method which is used in many fields for valuing subjective, introspective phenomena.

SG uses a probability concept and is offered as an iterative paired comparison (Torrance *et al.*, 1972; Gafni, 1994). Participants are presented with two alternatives, and asked to select the preferred one. We specified the first alternative as a description of the stable state after RT or S, while the other alternative offered a gamble with mutually exclusive probabilities for two reference outcomes, a good and a bad one. The reference outcomes in our study were perfect health with a probability (p) or death with probability ($1-p$). By varying the p-level of the uncertain outcome in steps of 5%, participants were able to indicate the point of indifference between the two alternatives (preference = p).

The TTO method uses a time concept. Participants trade off length of survival and health status. We have specified the first alternative as a description of RT or S with a duration of 10 years. The competing alternative offered perfect health of shorter duration. The point of indifference between the two alternatives was reached by varying the duration (y) in perfect health in steps of 0.5 years (preference = $y/10$). Detailed description of these two tradeoff techniques is formulated elsewhere (Torrance *et al.*, 1972; Weinstein *et al.*, 1980; Torrance, 1986).

The third method, VAS, required the participants to value a particular health state by placing a mark on a 100 mm. line with descriptive anchors at both end-points (Wewers & Lowe, 1990). In this study, "best imaginable health state" and "worst imaginable health state" were used as anchors for the VAS (preference = distance in mm. from "worst imaginable health state" to mark / 100).

Study design

All 24 participants were interviewed during three visits with two-week intervals in between. At the first visit, one complete valuation interview (I) was held. At the second visit two identical valuation interviews (II, III) were performed. In between these two valuation interviews detailed additional clinical information on T3 laryngeal cancer was offered. Both treatment modalities and their possible outcomes were thoroughly explained in layman terms. This part of the study was carried out by a trained interviewer who used several visual and verbal techniques to offer the information. During this visit the participants were free to ask questions about any topic. The fourth interview (IV), on the third and last visit, was identical to the previous valuation interviews.

The order of the valuation methods and the treatment outcomes was completely randomized over the subjects and the four interviews.

Analysis

To test concordance between the valuations on successive interviews, Intraclass Correlation Coefficients (ICCs) were computed. Contrary to Pearson correlation coefficients, ICCs include level-effects between different measurements. Therefore, ICCs will have lower values than the corresponding Pearson correlation coefficients. Analysis of variance for repeated measurements (MANOVA) was used to detect significant effects for the four experimental

factors (Interview, Group, Method, Treatment) and their interaction terms. Because of the small sample sizes, we labelled p-values under .10 as significant; p-values between .10 and .20 as a "tendency" towards difference, in order not to loose too much power in detecting underlying differences. MANOVA was also used to test three specific assumptions (profile analysis) (Tabachnick & Fidell, 1989). These analyses were performed separately for each group and for both treatments. The first assumption was: 1) Do the preferences for the three valuation methods differ by a constant level at each of the four interviews (e.g., parallel profiles)? If they differ, we may test the stronger assumption: 2) Are the preferences for the three methods equal at each of the four separate interviews, i.e., is the constant difference of (1) equal to zero (coincident profiles)? The third question that may be addressed if the two preceding hypotheses are not rejected is: 3) Are the preferences the same for all three valuation methods at all four interviews (coincident + level profiles; e.g., "flatness")?

Results

Participants' characteristics

The mean age of the former patients, 62.3, and the matched lay volunteers, 60.3, was almost identical, as intended. Males were in the majority, due to the specific kind of cancer we were studying. This predominance of males (n = 10) over females (n = 2) replicated in the group of the lay volunteers, namely seven males and five females.

Descriptives

The average preference for the outcome after radiation therapy (RT) was higher than for the outcome after surgery (S) for all three methods and all four interviews (Table 1). As expected, standard-gamble (SG) yielded the highest preferences, visual analogue scale (VAS) the lowest and the time-tradeoff (TTO) preferences were in between (Llewellyn-Thomas *et al.*, 1984a; Boyd *et al.*, 1990; Nord, 1992; Bass *et al.*, 1994). Regardless of the method and interview, standard deviations were systematically greater for S than for RT. Especially the standard deviations for S elicited at the first interview were rather large.

Difference scores (RT–S) were highest at the first interview, irrespective of the method used. The lowest difference scores were elicited at the third interview which was held at the same visit as the second interview, directly after the offering of additional information.

Stability of the preferences

Figure 1 shows the mean preferences for the three methods according to type of respondent and treatment. The means are not equal at the four interviews and the overall differences showed a tendency towards statistical significance (p = 0.13, Table 2). In particular, some substantial instabilities can be noticed. First, there is an increase in the valuation of the former patients between interview I and II for treatment S. Second, between interviews II and III we

Table 1

Means (standard deviations) of the utilities for the two treatments (radiation therapy, surgery) elicited on the four replication interviews. Subdivision according to the type of participants (former patients, lay volunteers) and valuation methods (standard-gamble, time-tradeoff, visual analogue scale). Between interview II and III additional clinical information was offered

Groups and Valuation Methods	Radiation Therapy				Surgery				Difference scores (Radiation Therapy - Surgery)			
	Interview				Interview				Interview			
	I	II	III	IV	I	II	III	IV	I	II	III	IV
Former Patients (n = 12)												
Standard-Gamble	.87 (.13)	.88 (.14)	.86 (.18)	.90 (.14)	.69 (.32)	.77 (.18)	.79 (.17)	.79 (.22)	.18 (.31)	.11 (.19)	.07 (.16)	.11 (.24)
Time-Tradeoff	.87 (.13)	.86 (.18)	.85 (.16)	.89 (.13)	.66 (.30)	.73 (.19)	.78 (.19)	.73 (.22)	.21 (.34)	.07 (.21)	.07 (.18)	.16 (.24)
Visual Analogue Scale	.80 (.11)	.83 (.13)	.83 (.11)	.77 (.20)	.50 (.31)	.60 (.18)	.63 (.18)	.55 (.27)	.30 (.35)	.23 (.36)	.20 (.28)	.22 (.39)
Lay Volunteers (n = 12)												
Standard-Gamble	.88 (.11)	.83 (.17)	.74 (.21)	.84 (.17)	.68 (.24)	.66 (.22)	.68 (.22)	.72 (.19)	.20 (.24)	.17 (.10)	.06 (.16)	.12 (.13)
Time-Tradeoff	.86 (.15)	.83 (.12)	.73 (.21)	.84 (.11)	.59 (.29)	.59 (.26)	.58 (.26)	.67 (.18)	.27 (.28)	.24 (.18)	.15 (.23)	.17 (.14)
Visual Analogue Scale	.73 (.15)	.78 (.07)	.71 (.17)	.74 (.20)	.43 (.20)	.55 (.24)	.49 (.24)	.51 (.23)	.20 (.23)	.23 (.17)	.22 (.23)	.23 (.19)

Table 2
Analysis of variance (MANOVA; repeated measurements) for the four experimental factors: the four replications of the assessments (Interview); the former patients versus the lay volunteers (Group); application of the three valuation methods standard-gamble, time-tradeoff, visual analogue scale (Method) and the two treatments, radiation therapy and surgery (Treatment)

Factor	p-Value
Interview	.13
Group	.14
Method	< .01
Treatment	< .01
Interview × Group	.07
Interview × Method	.17
Interview × Treatment	.03
Group × Method	.52
Group × Treatment	.72
Method × Treatment	.02
Interview × Group × Method	.80
Interview × Group × Treatment	.57
Interview × Method × Treatment	.93
Group × Method × Treatment	.45
Interview × Group × Method × Treatment	.48

* P-values between .10 and .20 are regarded as a "tendency" towards difference.
** P-values under .10 are indicated as "significant".

see a reduction of the preferences for RT by the lay volunteers. Third, at the last interview the preferences of the former patients and the lay volunteers, just as at the first interview, have become very similar, due to changes between interviews III and IV. The most significant interaction term related to the stability of the preferences was Interview × Treatment ($p = 0.03$). This effect was on account of a systematically higher preference for S at the second interview, especially for the former patients (details, see below). The interaction term Interview × Group was also significant ($p = 0.07$).

Preferences were most stable for S, both for the former patients and for the lay volunteers (Table 3). With the exception of one low reproducibility coefficient (ICC = 0.19), all coefficients for S were > 0.60. For RT, the reproducibility coefficients were much lower, especially for the group of lay volunteers. Examining the three methods separately showed that in particular the preferences elicited by SG were stable between the successive interviews. This is illustrated by the ICC-interquartile ranges for each method. The SG method showed the highest range (0.67 – 0.83), the TTO method somewhat lower (0.57 – 0.77), whereas the VAS method showed the lowest range (0.47 – 0.67).

The ICCs also make it clear that the means in Figure 1 have to be carefully interpreted. If two means at succeeding interviews are almost equal, the ICC can nevertheless be low due to

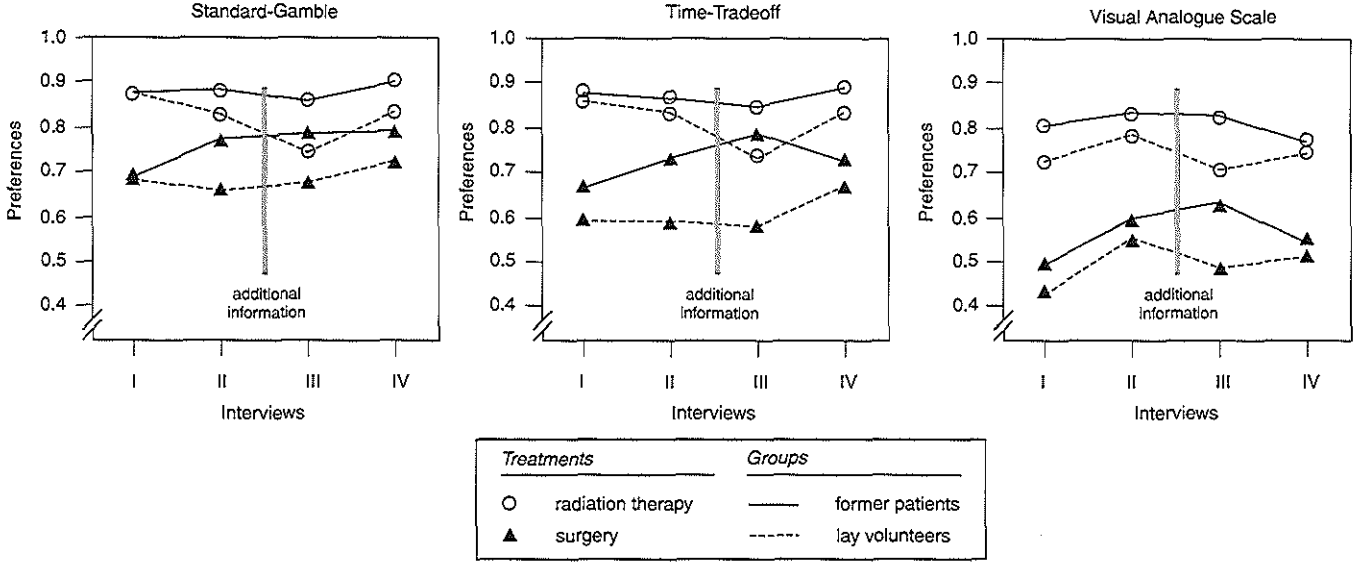


Figure 1
Mean preferences for the two treatments (radiation therapy, surgery) subdivided by the two groups of participants (former patients, lay volunteers) held at four interviews (I – IV) using three valuation methods (standard-gamble, time-tradeoff, visual analogue scale). See also Table 1

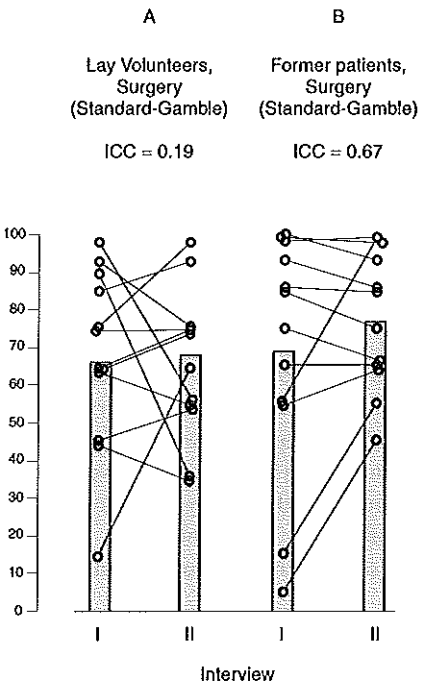


Figure 2
Detailed examination of two intraclass correlation coefficients depicting the two times 12 individual preferences (nodes) for surgery as measured with the standard-gamble method at valuation interviews I and II (bar indicates mean preference)

"massive" individual preference shifts within a group (Figure 2a). On the other, even reasonably high ICCs do not preclude a small difference between preference means together with a considerable shift of preferences for some of the individuals (Figure 2b).

Comparability of the valuation methods

As expected the SG method and the TTO method gave comparable results. The mean VAS valuations for S of the former patients were substantially lower than for the other two methods. Accordingly the analysis of variance (Table 3) revealed a significant difference ($p < 0.01$) between the valuation methods. One interaction comprising the term Method was significant: Method × Treatment ($p = 0.02$). This is undoubtedly caused by the systematically lower preferences for S as elicited by the VAS.

Table 4 gives results of the profile analysis which tests the comparability of the methods in more detail. The test of parallelism for the three methods holds quite well. Generally, the preferences elicited by the three methods showed a constant difference across the four interviews. The subsequent assumption, e.g. the coincidence of the methods, appeared to hold

well for RT but not for S, with two borderline significant p-values. Thus, for RT the methods roughly provided identical preferences, although their means may still differ across the interviews only. The most selective assumption, e.g., equal means for the three methods at all four interviews, was not rejected for the RT preferences of the former patients ($p = 0.90$).

The two groups of respondents

The finding that on average RT was valued higher than S, irrespective of the group of respondents (Table 1), was significant ($p < 0.01$, Table 2). Moreover, the interaction term Group × Treatment was not significant. Surprisingly, although Figure 1 suggests that former patients produced valuations that were systematically higher than lay volunteers, no clear-cut statistical significance between the two groups ($p = 0.14$) was detected, only a tendency. We believe that this can be attributed to: the interaction of Interview × Group and the relatively high dispersion of the individual valuations (as well as the already mentioned "massive" individual preference shifts within a group between successive interviews).

Another aspect that can be noticed in the context of the two groups of respondents is that the former patients valued S differently at the interviews II – IV compared to the first interview. For the lay volunteers, such a general difference could only be detected for the VAS method.

Table 3

Reproducibility of preferences for three successive retests of the interview (I – II, II – III, III – IV) as measured by intraclass correlation coefficients (ICC), for he three valuation methods (standard-gamble, SG; time-tradeoff, TTO; visual analogue scale, VAS), the two treatments (radiation therapy, surgery), and the two groups of participants (former patients, lay volunteers).

Groups and Treatments	Interviews I and II			Interviews II and III			Interviews III and IV		
	SG	TTO	VAS	SG	TTO	VAS	SG	TTO	VAS
Former Patients (n = 12)									
Radiation Therapy	.72	.66	.23	.91	.75	.85	.75	.39	.46
Surgery	.67	.75	.63	.93	.92	.66	.70	.84	.67
Lay Volunteers (n = 12)									
Radiation Therapy	.46	.21	.56	.74	.57	.47	.56	.41	.20
Surgery	.19	.72	.61	.81	.93	.82	.83	.77	.73

Table 4
Concordance of the three valuation methods (standard-gamble, time-tradeoff, visual analogue scale) tested by MANOVA (Profile Analysis); the table gives the p-values for the tests of parallelism (1), coincidence (2) and flatness (3). Subdivision according to group (former patients, lay volunteers) and treatment (radiation therapy, surgery).

Groups and Treatments	Tested Assumptions		
	Parallelism (1)	Coincidence (2)	Flatness (3)
Former Patients (n = 12)			
Radiation Therapy	0.71	0.35	0.90
Surgery	0.85	0.09**	0.02**
Lay Volunteers (n = 12)			
Radiation Therapy	0.68	0.23	0.02**
Surgery	0.35	0.08**	0.15*

** Assumption statistically rejected if p-values under .10.
* P-values between .10 and .20 are regarded as a "tendency" towards statistical significance.

Values also reflect the relative importance of the effect

Note: assumption (3) may be tested if assumptions (2) and (1) hold;
assumption (2) may be tested if assumption (1) holds.

Discussion

In this experiment, we studied the stability of preferences in order to analyze the validity of the practice of "one-shot" elicitation of preferences. Such preferences for treatments are often merged with prognostic data (life expectancies, probabilities) in clinical decision analysis (decision tree analysis) and also in cost-effectiveness analysis (of better: cost-utility analysis) (Torrance & Feeny, 1989). In this study, in order to investigate the stability of preferences, we studied the repeated elicitation of treatment preferences. We used a balanced experimental design, in order to be able to detect other factors that may contribute to preferences. In addition, apart from the two treatment options and the replication of their valuation, we investigated the comparability of three elicitation methods and the similarity between two different groups of respondents.

A stable effect was detected for the order of preference of the two treatment modalities. The health state after radiation therapy (RT) was on average valued higher than the health state after surgery (S) under all conditions by both the former laryngeal cancer patients and the lay volunteers (in clinical applications of the formal analytical approach, this treatment effect will be the only factor of interest). Preferences elicited by the lay volunteers tended to be lower than those of the former patients.

Preferences were not stable across at the four valuation interviews. Because the four preference measurements fell within a short time period, in which the health status of the participants can be assumed to be stable, these variations can, in our opinion, be attributed to

measurement error and the cognitive evaluation of the clinical information. The offering of more explicit information is probably responsible for the changes along interviews II and IV. Because interviews II and III were at the same visit, the participants may have remembered their preferences at interview II and for this reason adhered closely to these preferences at the subsequent interview, despite possible mental shifts due to the additional information. Thus it could be speculated that the change in preference between interviews III (visit 2) and IV (visit 3) may (partly) be attributed to the delayed influence of the additional information, i.e., after an interval of two weeks the participants may no longer remember their preceding preferences, but may have fully utilized the additional information offered. With respect to the stability of the individual valuations at the succeeding interviews, no clear differences nor systematic resemblance could be observed between the three methods.

As expected, standard-gamble (SG) preferences were highest and visual analogue scale (VAS) preferences lowest. We should recognize, however, that differences in level between methods need not to be a problem for subsequent decision analysis as long as the relative differences between health states are preserved. After all, the main purpose is to elicit preferences for treatment outcomes that represent relative preferences for the treatment outcomes. Another conclusion related to the methods is that irrespective of the group of participants and the interviews, both the absolute and the relative difference between the means of RT and S was smallest for SG and largest for VAS. The VAS method in this study comprised a lower anchor-point (worst imaginable health state) that was different to the trade off methods (death). Although this should have theoretically compressed the range of preferences on this scale somewhat, the VAS method showed larger differences between the two treatments. The expected correspondence between the two tradeoff methods, SG and the time-tradeoff (TTO), was largely confirmed.

An important conclusion arising from this study was that apparently, lay volunteers without any experience of the objectives of the study, seemed on average to arrive at the same preferences as former patients after repetition of the valuation task. It may also be observed that almost identical preferences were observed at the first interview. However, it should be noted that on the first valuation the variation of the assessments was much greater for S. Probably more important is that the preferences for S were definitely altered at the last interview in comparison with the first one. Another difference between the first and the last valuation interview is that the differences between the two treatments were smaller at the last interview, irrespective of the valuation method used or the group of participants.

Repetition of the valuation task and the offering of (background) information seems essential to obtain valid assessments. A general population sample appeared to make assessments that were comparable to those obtained from experienced subjects. Furthermore, sufficient time intervals seem necessary for subjects to process the offered information. This finding may be interpreted as evidence for the validity of using assessments of layman to acquire valid preferences from the societal perspective. However, at the individual (patient) level much more cautious conclusions should be drawn. Statistical strategies have been used to estimate the average reliability of individual preferences, but no definite conclusions can be made considering specific individuals (Roest *et al.*, 1997).

Because clinical decision science and utility theory do not assume either an aggregated perspective or repeated decisions, clinical decision analysis is assumed to be applicable for one-shot individual medical decisions. Our results indicate that conclusions can be questionable when preferences are obtained from a one-session valuation task. The preferences obtained at the first interview proved to be very unstable both at group level and at individual level. Replication of the rather complex valuation tasks and the rehearsal of the offered information seems necessary to elicit valid preferences. For clinical decision making, at least in this study, SG seems the most reliable method, whereas for obtaining preferences from the societal perspective (mean preferences), all three methods seem to be equally suitable. Moreover, our study suggests that these latter assessments may be validly attained from a sample of the general population. New studies on this subject should in our opinion be carried out with rigorous design like ours. Albeit that, when possible, such studies should be conducted with more participants and preferably among patients, who are closer to the actual decision than in the present study.

Acknowledgments

This study was supported by a grant from the Dutch Cancer Society (grant DDHK 91-05).

We express our gratitude to the Dutch Cooperative Head and Neck Oncology Group, Utrecht, the Netherlands, for enabling us to use some of their data on the outcome of T3 laryngeal cancer treated by either radiation therapy of surgery.

Chapter 10

Discussion and Conclusions

Discussion and Conclusions

The central premise of this thesis is that it is possible to quantify health and potential health gains and losses within the context of decision support. A precondition of realizing this goal is the elicitation of reliable valuations for health outcomes. However, unlike the measurement of survival, the procedures used to obtain valid and reliable values for health status are not simple and straightforward. The same applies a fortiori to the procedures to combine such values with survival data to produce summary health values like Quality-Adjusted Life Years (QALYs).

In the following three sections, we will discuss the results of the research reported in Chapters 2 – 9. First, the QALY approach in general is discussed. Both the theoretical points raised in this thesis and the empirical results emanating from the QALY approach are addressed. Next, a detailed discussion follows about the valuation of one of the two basic elements of the QALY model, namely health status. Third, we discuss the findings of the studies from both the individual (clinical decision making) and the societal perspective. We conclude with a brief discussion about the future perspective for quantifying health outcomes.

The QALY approach

Regardless of its operationalization, the validity of the general QALY approach has been criticized. Probably the main objection concerns the implicit ethical and philosophical consequences of its application (Harris, 1987; Smith, 1987; Harris, 1988; La Puma & Lawlor, 1990; Selai & Rosser, 1993; van der Wilt, 1994; Cohen, 1996; Williams, 1996). In particular the distribution properties (justice, equity) of QALY measures in relation to age and income are questioned (Lee, 1967; Carr-Hill, 1989; van Busschbach *et al.*, 1993; Nord *et al.*, 1995; Dolan *et al.*, 1996). Two types of assumed distribution "distortions" can be noticed: 1) distortions due to different appraisals of health outcomes (e.g., different health perception at older age, different health perception as a result of poverty), 2) distortions attributable to the algorithm of the QALY approach, due to different (expected) life years at different ages. From a scientific point of view it should be agreed that *any* quantitative outcome measure ultimately discriminates, as such measures aim to select people on the basis of measurable benefit. Hence, criticism is not directed specifically at the QALY but at any pathophysiological or clinical measure of outcome. The issue cannot be resolved easily however as personal characteristics (age, ethnicity, sex, social-economic class) determining effectiveness are often the same as those discussed in the context of discrimination (Bonsel, 1991).

So far, the evidence supporting the view that background characteristics of respondents significantly influence assessments, is weak. It seems unlikely that any of the reported small (linear) effects of age or education will affect choices between medical interventions. We conclude therefore, that there are no specific objections to the QALY approach from the distributional point of view.

The apparent drive to apply QALYs at the individual and societal level can only be explained by an underlying implicit acceptance of the soundness of the concept of combining

length of life and quality of health in one overall measure. Despite the broad application of the QALY model, several issues related to the QALY approach require further elaboration.

- The theoretical foundation of the QALY approach is not developed well enough to legitimate the application of QALY measures in general.
- The approach is usually applied under very restrictive assumptions.
- Little focus has been directed at length of life (L) and its subjective valuation, $U(L)$, or its role in more aggregate QALY models.
- A variety of strategies and methods is used to value the elements of the QALY model, in particular the assessment of the quality of health (Q), e.g., health status.
- At the level of clinical studies, little knowledge is available about the distribution properties of the resulting composite QALY measures, which hampers, among other things, statistical testing and sample size calculation. This may be due partially to the limited research focused on the prevalence of, e.g., QALY-ples or complete health profiles (Q, L).
- Until now the QALY approach has not been incorporated in Randomized Clinical Trials in a standardized way, whereas applications at the level of individual decision making have been restricted to experimental conditions.
- The QALY approach is not applicable for all health situations. For instance, assessments in the field of fertility care and care for the (already) mentally handicapped are not valid, because health care is in these cases not primarily directed to health improvement per se.

The element of time in QALY models

Here we will elaborate on one general issue, namely, the life years/survival element in QALY models. Survival (L) is a common and important measure to express benefits of medical interventions. It also is one of the key elements in QALY models. As already mentioned, most QALY modeling has hitherto been based on the additive variant of the QALY model. In our terminology this is referred to as the Partially Decomposed QALY Model (Chapter 2) for non-constant health states. Modeling based on this model has in almost all studies been performed without the distinct and separate valuation of L (either $U(L)$, $U(Q, L)$, $U(Q, L)$ etc.). As a result, little is known of the contribution of L in the QALY models from an empirical point of view (Stalmeier *et al.*, 1996). Conventional QALY modeling is carried out by attributing the same weight to each life year, i.e., each life year is treated as one standardized, say, physical unit. This is a simplification and probably does not conform to the concept of time as experienced and valued by subjects. Studies on the appraisal of survival in relation to health are rare. This may be explained by the paradigm of Cost-Utility Analysis, where it is generally assumed that the discount rates for costs (i) and health effects (r) have to be the same (see for a detailed discussion about this subject: Gold *et al.*, 1996). One of the studies directed to this topic evaluated the L parameter in relation to the subjective discount rate of life years (Chapman & Elstein, 1995).

When referring to the survival element in the context of the QALY approach, we do not mean the estimation of the survival parameter in clinical research, but the subjective valuation of the time element, e.g., $U(L)$, assuming that L itself is properly measured. We expect $U(L)$ not to be a simple linear or exponential function of L, whereas experience up until now does not support the assumption of independence between $U(L)$ and $U(Q)$.

When statements about the duration of health states are introduced into the stimulus component of the valuation task, the elicited values seem to be affected. Sutherland *et al.* (1982) reported empirical evidence that attitudes of health professionals towards survival depended both on the amount of time to be spent in a hypothetical bad health state (followed by death), and on the quality of the state. Subjects apparently make such assessments using a subjective variable, the concept of "maximal endurable time" in a given state. When this time was exceeded, attitudes toward additional increments of survival changed. The longer a bad health state lasts, the more intolerable it becomes and the lower the valuation assigned to that state. Again this points to the need for assessment of more aggregated outcomes (QALY Models III and IV).

In our study, we did not investigate *U(L)* empirically, but we did focus on the main assumption within the context of the basic additive model, that is that the value of 8 years in health state A followed by 2 years in health state B is equal to the value of the same states in reverse order (Chapter 5). We observed a small but distinct sequence effect that could not be explained by discounting. This effect was too small to disclaim the model, but too large to ignore the effect. According to our knowledge, this is the first empirical evidence and quantification of the sequence effect. We conclude that both the maximal endurable time effect and our sequence effect, e.g., $U[(Q, L)_A, (Q, L)_B] \neq U[(Q, L)_B, (Q, L)_A]$ (Figure 1), essentially are the same violations of the additive model (assuming that both effects are adjusted for time preference effects).

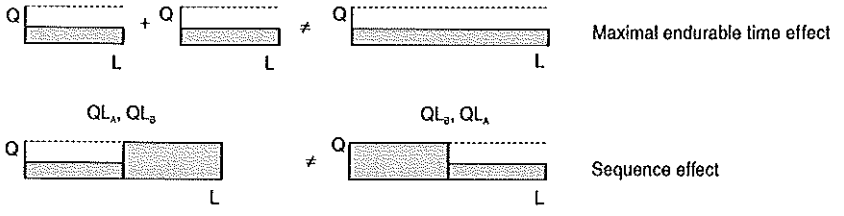


Figure 1
Visual representation of the maximal endurable time effect (above) and the sequence effect (below), both effects adjusted for time preference effects

Adaptations of the QALY model

Several proposals have been made to correct or adjust QALYs. Here, "correct" refers to a situation where the value of summary measures expressed in QALYs is thought to be intrinsically biased, e.g., by the use of a valuation method. "Adjusting" a QALY refers here to a situation where the resulting measure in itself is unbiased, but still for some purpose needs adaptation. Two most prominent cases in the latter category are discounting of life years and age-adjustments. Both are adjustments that can be dealt with as part of a valuation task itself or afterwards by analytical strategies. Discounting is a conventional procedure in financial and

economic calculation. It implies that future gains (expressed in monetary units) are valued less compared to the same gains achieved immediately (Drummond *et al.*, 1997). As already mentioned, in cost-utility analysis, it is generally assumed that both the numerator (costs) and the denominator (QALYs) of the cost-utility ratio are subject to the same discounting rates (Gold *et al.*, 1996). Normally, discount rates of 3% or 5% are used for this purpose. The question of whether the discounting phenomenon for health outcomes is identical to the discounting of monetary goods has been investigated and empirical studies have shown that under various conditions this assumption does not hold (Redelmeier & Heller, 1993; Chapman & Elstein, 1995; Krabbe & Bonsel, 1998). Apparently, the effect of the time element in the evaluation of health is more complicated and diverse than for costs. The straightforward use of constant discounting rates that may be appropriate for costs seems to be much less justifiable for health outcomes.

Special consideration is required for the effects related to age. First, the valuation of health states, but also one's own health, may depend on the age of a person. For example, a health state comprising minor complaints may be differently assessed by a 30-year-old than by an 80-year-old as the reference "average" health condition is not the same for the two ages. A similar argument is valid for age in case L is valued, instead of assuming $U(L) = L$. Second, treatment of a disease by a medical intervention is often more effective for younger people, because unspecific life expectancy is greater for younger people. Third, age-weighting is proposed and discussed by several researchers (Murray & Lopez, 1994; Barendregt *et al.*, 1996; Williams, 1997) for reasons of justice, for which economic and social arguments are used. For instance, Murray *et al.* (1994) have used an age-weighting function in their Global Burden of Disease study at the societal level. Stating that in all societies, social roles vary with age, they considered it appropriate to value the time lived at particular ages unequally. In their model, life years gained at young age and older age were valued less than the life years gained at middle age (age 9 – 54). To a certain extent, discrimination by age has an intuitive appeal. It also seems to reflect the unequal valuation of life at different stages. This is often observed, for example, in developing countries and earlier civilizations, whose inhabitants sometimes appear to display relative indifference towards the mortality of young children compared, to say, the death of young adults. However, it is far from easy to define the appropriate function to arrive at a suitable measure for acceptable inequity (Barendregt *et al.*, 1996).

If deemed necessary, both adjustments (discounting of health outcomes and age-weighting) are best dealt with *after* the (separate) assessment of the elements of the QALY model being used. It seems essential to gather empirical evidence on this topic in order to overcome the lasting discussions among theorists and to achieve some consensus on this issue (see also: Holmes, 1995).

The valuation of health status

Within the five different stages that we have distinguished in Chapter 2 in order to arrive at summary values for health, the composition and quantification of health states takes place at respectively Stage II, where the construction and conceptualization of the scenarios for the health states takes place, and at Stage III where the valuation of these scenarios is carried out. Here the choices are: 1) the use of naturalistic or multiattribute scenarios and some related

issues, 2) the valuation method, 3) the composition of respondent panels, 4) the stimulus mode, 5) the response mode and 6) the analysis approaches. Choices 1 – 3 and 5 – 6 have been studied in this thesis and will, except for aspect 5, be discussed separately in the following sections.

Scenarios

Both the naturalistic and the multiattribute way of framing health-state scenarios were studied in this thesis. In Chapter 3 – 6, all experiments were based on multiattribute scenarios constructed by the EuroQol health-status classification system. Valuations of these scenarios where studied within the context of the Partially Decomposed QALY Model. In contrast, Chapters 7 – 9 present results of valuation tasks carried out by the use of two variants of naturalistic scenarios of health status only, and of lifetime health profiles that include prognostic elements. The experiments in Chapter 3 – 6 were primarily focused on the psychometric properties of the valuation methods (validity, reliability) by valuing an arbitrary set of health states, while in Chapters 7 – 9 the application of clinical decision analysis was the primary focus. This perspective of application determined the health states c.q. health profiles to be valued. The main interest was not in the valuation methodology as such but in the interpretation of the individual responses. We did not compare naturalistic scenarios directly with multiattribute scenarios, as has been carried out recently by others (Stouthard *et al.*, 1997).

From our experiments, we conclude the following. In clinical decision analysis, naturalistic scenarios seem more feasible to the extent that the number of scenarios to be assessed is usually low. Additionally, naturalistic scenarios are more appropriate if we want *involved* respondents (patients) to understand the information offered by scenarios, without additional instruction. Furthermore, the use of naturalistic scenarios usually avoids the so-called dominance of scenarios, that is the presence of a scenario that is consequently superior according to all the aspects described (see: Chapter 4).

The drawbacks of naturalistic descriptions are the following. First, the so-called labeling effect may be present. If respondents have to value a scenario that is associated with or explicitly labelled as, e.g., AIDS or cancer, the valuation probably will be lower compared to a similar health state with unknown etiology. In other words: the additional information covered by a label is liable to distortion by negative or positive connotations, particularly if the panel lacks medical education. Second, if non-involved respondents are used, naturalistic scenarios bear the risk of not being understood, particularly if descriptions are based – as they usually are – on medical concepts and terms. The same may happen for involved respondents, because we cannot safely assume that these respondents will have, by definition, a full understanding and knowledge of all the aspects and consequences related to their own disease. The heterogeneity of a panel response may therefore reflect a different perception of the stimuli rather than a different response to the same stimuli. Third, no value function across the, essentially, nominal stimuli can be estimated, which implies that all the relevant health states have to be separately valued and consistency checks are difficult, if not impossible to carry out. Fourth, it is difficult to translate clinical data (frequently also comprising prognostic and duration health-elements) into naturalistic descriptions in a reliable and valid (replicable)

way, compatible with the rather straightforward use of, e.g., descriptive EuroQol data of patients in multiattribute scenarios.

It can be argued that multiattribute scenarios are also arbitrary and depend on the choice of existing health-status classification systems. However, the few systems available, that is the Rosser & Kind index (1978), the 15D developed by Sintonen and colleagues in Finland (1981; Gold *et al.*, 1996), EuroQol (Brooks, 1996), Health Utility Index Mark-II (Torrance *et al.*, 1996) and the Quality of Well-Being scale (Kaplan & Anderson, 1988), mainly differ in refinement rather than in scope or comprehensiveness of the health-status concept. The EuroQol classification system that we have used (Chapter 3 – 6) seems a sensible choice, although an extension, for instance by adding a cognition dimension (Chapter 6), may be recommended, as well as an increase from three to four or five levels (Bonsel & van Agt, 1995; Selai, 1996).

For the naturalistic scenarios as well as for the multiattribute scenarios, the indication of time-elements is a crucial factor. Two distinct time-elements must be specified in scenarios: the *duration* of the state and the (expected) *end-point*. In the near future studies are needed that compare results of single QALY-ples, episodic health profiles and lifetime health profiles obtained with both types of scenarios. Amongst others, the additive utility independence assumption can then be tested in detail. Related to the scenarios, some other stimuli related factors are important. Kahneman and Tversky (1982) have reported different (risk) attitudes of subjects towards the assessment of equivalent scenarios described respectively in terms of saving or losing of life. This "framing" of scenarios, for instance, stating treatment outcomes (health states) as gains or losses, probably will affect the valuation. The same applies to the exact phrasing of accompanying statements in the operationalization of the valuation task of a valuation method.

Other design aspects are, e.g., the order of presentation and the use (or not) of interviewer support. In our experiments, different degrees of support were present: from unsupported postal questionnaires (Chapter 6), and group instructions (Chapters 3 – 5) to face-to-face interviews (Chapters 7 – 9). In the latter case, we advocate a high degree of standardization. This can be achieved by training and the use of a script.

Valuation methods

Which valuation method is the most appropriate is not self-evident (Bonsel, 1991; Nord, 1992; Essink-Bot, 1995). Each valuation method has its own specific methodological consequences and there is still no consensus about which valuation method should be used. The decision about which method to use in a particular study is, as for descriptive health-status measures (Essink-Bot *et al.*, 1997), based on scientific and, it appears, extra scientific considerations. Arguments to apply tradeoff methods seem valid but still we must be aware that with the exception of the Healthy Years Equivalents (HYEs) introduced by Mehrez and Gafni (Mehrez & Gafni, 1989; Gafni, 1996), it is only health status (Q), not Q and survival (L), that has been valued. Thus superiority of the conventional tradeoff methods can not be substantiated.

Besides the choice of one valuation method in favor of others, another relevant choice that has be made irrespective of which valuation method is to be used, is the choice of the anchor-states. It should be kept in mind that despite the convention to use death as the lower

anchor point, methods such as standard-gamble and time-tradeoff can be operationalized equally validly (and perhaps conceptually more consistently) using other health states as lower anchor point. The same applies for upper anchor points. As long as health outcomes can be positioned between the two anchor points, any anchor point may serve. If anchor points other than perfect health and/or death are used, the utilities obtained using such non-standard operationalization are different and need rescaling factors for inter-study comparisons (see: Chapter 3 and Fryback & Lawrence, 1997). However, we encountered considerable difficulty in convincing the reviewers of the journal in which Chapter 3 was published of the fact that the use of a specific anchor point ("death" rather than "worst imaginable health state" in Chapter 3) was neither prohibitive, nor an essential property of these two methods. The reluctance of the reviewers is probably due to the general convention of using death as the lower anchor point and the lack of a theoretical QALY model.

Several studies have been performed in which different health states were valued, including "death". Subsequently, the elicited values were transformed by setting the individual value of death at zero. As a consequence of this approach, health states valued as worse than death are assigned as negative values. It is apparent from Appendix A that negative *Qs* are inconsistent. Mainly for this reason, death cannot be expressed as a special variant of *Q*. All we can say about death is that it is a concept that can be cognitively mapped to some extent as a bad health state with an undefined *L*. Therefore, death as the lower anchor point is conceptually different from anchors expressed in terms of lasting health states, despite "death" being popular as the lower anchoring label in valuation tasks.

Standard-gamble (SG) is the classical method for the elicitation of utilities (values) in medical settings (individual perspective) and has occasionally been proposed as the "gold standard" (Froberg & Kane, 1989b; Gafni, 1994; Gold *et al.*, 1996). SG is the only valuation method that is operationalized with probabilities (uncertainty c.q. risk) as the standard component of the tradeoff task. If a specific person behaves in agreement with Expected Utility Theory (EUT), then it is postulated that the SG method provides utilities. Richardson (1994), however, has pointed out that their is a clear difference between, on one hand the risk situation (p) created by the use of SG method itself, and on the other hand the empirical probabilities (p^*) that are involved with the outcomes of the medical intervention. Only in some special cases will the probabilities of the stimulus (outcome medical intervention) and the probabilities of the response (SG) coincide. For studies at the societal level, uncertainty is essentially not an issue. Moreover, the SG method is difficult to understand, especially by non-trained respondents, as is apparent from the necessity to use a visual (analogue) device ("pie", chance board) to reflect probabilities. For these reasons, the superiority of SG, both on conceptual and on empirical grounds, can be doubted.

Time-tradeoff (TTO) has been presented as a valuation method closely related to the SG method, but easier to perform and to understand. In our studies we experienced that both students and former patients, indicated that they found the TTO method less complicated in comparison to the SG method. However, there is one serious effect that has to be dealt with when using the TTO method, which is the inherent tradeoff of time which takes place due to the operationalization of the TTO task by the time element (L). This aspect has usually been ignored.

Another limitation of the conventional operationalization of TTO and SG with perfect health and death as anchor points, is that both methods do not discriminate between relatively good health states, which basically can be attributed to two features: ceiling effects and calibration properties. Ceiling is the presence of non-equidistancy (here: compression) between scoring points, given some gold standard. Calibration (or degree of granularity) is the density of scoring points given some gold standard. Ceiling effects occur on the account of the inherent task of these two methods, i.e., the sacrificing of probability of good outcome (SG) and life years (TTO). Subjects are reluctant to sacrifice one of these entities to improve mild health states. In the end this is a matter of normative judgment. If the reluctance to tradeoff for mild health states reflects a societally supported attitude in competing choices under (budget) restrictions, then tradeoff methods are valid and appropriate, rather than unreliable and insensitive. Calibration aspects are also induced by the operationalization of SG and TTO, for it is apparently difficult for people to express a difference between, for instance, a change of 2% versus 3% or between 2 life years versus two life years and one week. Tversky and Kahneman (1974; 1981; 1992) have demonstrated another distortion, showing that people are not interpreting probabilities conform algebraic, metric properties. This effect may introduce an additional bias to the SG method. Recently, techniques have been proposed (Wakker & Stiggelbout, 1995) to overcome – at least partially – this cognitive distortion of probabilities as calibrator when valuing *Q*. The validation of the soundness of this technique has yet to be established.

Our results confirmed the empirical equivalence of the SG and the TTO methods under optimal testing conditions (Chapter 3 and 4). Clearly these two tradeoff methods evoke approximately similar cognitive processing by people. Contrary to this, the rating scale (RS) and the visual analogue scale (VAS) do yield different values for health state valuation. Due to the fact that the RS is not based on a tradeoff concept, values elicited by this method are systematically lower (Torrance, 1976; Bass *et al.*, 1994; Bosch & Hunink, 1996; Krabbe *et al.*, 1997). Regardless of the different value functions associated with these valuation methods, our study showed that the RS is the method for which the internal consistency of the valuations of health states (reliability) among the respondents was highest (0.77). Reliability for the TTO was 0.65 and for the SG even lower, namely 0.58. In an earlier postal survey, which was also conducted using EuroQol health-state scenarios, RS reliability coefficients in the range 0.77 – 0.84 were observed (Essink-Bot *et al.*, 1993). Both results confirm the relatively good properties of RSs with regard to the reliability of the responses. An interesting result was the comparability between the RS method and the SG and TTO methods. In our study, valuations based on RS were distinct from, but strongly related to, values derived from the two tradeoff methods. A simple one-parameter power function sufficed to transform RS values to SG or TTO. If this relationship between RS and SG/TTO is not an artefact and can be confirmed in prospective studies, the valid use of the RS method combined with simple transformation, is a contribution to the feasibility for studies utilized at the individual and societal level.

A quite different approach is willingness-to-pay (WTP), which seems difficult to operationalize in prosperous Western countries. Western societies do accept care options which per capita by far exceed personal income, provided that associated risks are very small (which usually is the case). Moreover, many studies for which WTP is an issue are dealing

with, per capita, expensive care options. Within the framework of WTP, these costs make little sense to a Western respondent, neither as direct payment (sum of money is *above* imagination), nor as premium (risk is *below* imagination; see: Chapter 4). To this should be added the reluctance of laymen to participate in WTP studies in these countries where standard medical insurance covers the total population. As a result responses – if any – tend to be given in a very global way. Finally, problems due to response-modification by personal income can most likely be expected. When we compared WTP with other valuation methods, we were able to control for the effect of income, due to the specific sampling of the respondents. Nevertheless this method yielded values that were too global, and which are not comparable to the other methods. The reliability of health-state values elicited by this method were the lowest (0.49). In the past, a few studies have been performed using WTP (Thompson *et al.*, 1982; Thompson *et al.*, 1984; Thompson, 1986). Mainly because of the operationalization complexities of this method, there has been a reluctance to perform further studies, although, recently there seems to be a revival (Gafni, 1991; O'Brien & Viramontes, 1994; Chestnut *et al.*, 1996; Kartman *et al.*, 1996; O'Brien & Gafni, 1996; Stålhammar, 1996). Frequently these studies use a response modality which mirrors the clinical problem, in our words, are a copy of the research problem, e.g., the method is operationalized as a template (see: p. 173). Note, for example, that for the hypertension treatment case of Johannesson *et al.* (1991) patients were asked to indicate the highest amount of users fees they were prepared to pay (see also: Chestnut *et al.*, 1996). Such WTP-operationalizations, basically represented as an insurance problem, are based on the same questionable assumption underlying the combination of SG with surgery (risk situations) and TTO with chronic states (see: Chapter 2, p. 34).

The recently re-introduced person-tradeoff (PTO) method and its extended variant (see: Chapter 2, p. 36) was not included in this thesis, but some preliminary remarks based on experience so far can be made. Conceptually PTO is not a pure tradeoff method, because the task of this method is not trading off a particular health-related entity, but merely an equivalence task, i.e., how many (y) healthy people living for *one year* are equivalent to x people in health state i for *one year* (either x or y specified). Hence, the "tradeoff" does not correspond to an aspect that forms part of the health state itself. For this reason, it may lack the assumed advantages attributed to the established tradeoff methods. Extended PTO is constructed by combining the specific characteristics of two valuation methods: the equivalence technique and the tradeoff of life years. The complexity of extended PTO may be responsible for one or even more effects contributable to the valuation task of this method. Much of the assumed advantages of simple PTO and the extended version, in particular its assumed consideration for equity, are not confirmed empirically in studies published so far. Furthermore, it may be questioned whether the operationalization of the stimulus, e.g., x subjects dead after one year, is not merely an artefact similar to the artificial "one year health-status" stimuli in the Fully Decomposed and the Partially Decomposed QALY Models. Studies solely focused on the characteristics of the PTO, preferably in comparison with other conventional valuation methods under controlled conditions, are urgently needed.

Several textbooks (Kruskal & Wish, 1978; David, 1988; Streiner & Norman, 1995) claim indirect scaling perfectly suits the assessment of subjective phenomena and some papers (Fanshel & Bush, 1970; Hadorn *et al.*, 1992) suggest indirect scaling also to be promising for

health states. We discovered however (Chapter 4) that paired comparisons (PC) is not a useful valuation methodology for indirect scaling of health states that are derived from a classification system (i.e., multiattribute scenarios). This is fully attributable to the natural dominance that exists between many of the health-state scenarios, whereas PC is especially developed for scaling unconcatenate (by definition non-dominant) subjective stimuli. Its suitability for naturalistic scenarios remains to be investigated.

As we have already briefly discussed, response-biases are often introduced by configuring valuation tasks in a manner that the response (the calibrator) resembles/mimics the clinical problem in reality (stylized as stimulus). For example, a SG response mode resembles the direct, short term risk of a surgical procedure. TTO resembles the chronic health condition that can be improved, though at the consequence of less life years. The assumptions are that this resemblance adds to valid assessments. We will indicate this apparent conceptual error as the *template error*, an error which can be observed for all tradeoff methods. Where Richardson (1994) has mentioned the imperfection between the calibrator of the SG method (p) and the empirical probability (p^*) of the risk of an intervention under study, comparable stimulus/response substitution exist for the other tradeoff methods.

The TTO method introduces the phenomenon of *double discounting*. This refers to the phenomenon that time preference is simultaneously accounted for or plays a role in the valuation task (implicitly), and in the aggregative operations (explicitly). The exact nature of double discounting depends on the QALY model adopted. In case of the Fully Decomposed QALY Model and the Partially Decomposed QALY Model "time" is present in the response ("time" is the calibrator to arrive at a value), and as the life years element of the subsequent QALY calculations. The latter is discounted explicitly. However, the respondents may discount implicitly the response period, usually ten or 20 years, that may affect the values of the set A or Q derived from the task. Note, however, that this only affects the internal validity of TTO (see also: Appendix 5.1 of Chapter 5). The ultimate consequence is that health states (set A or Q only) receive slightly different values depending on the standard length of the time period to be traded off.

Genuine risks of double discounting occur in more aggregated QALY models, i.e., were TTO is applied to a stimulus including duration (the Single Profile QALY Model and the Multiple Profiles QALY Model). Here the value assigned to a QALY-ple or lifetime profile, may reflect discounting of the duration in the stimulus. Discounting effects may be present in the use of the response period, without serious problems. However, if in the Single Profile QALY Model, the values of QALY-ples are added, conventional discounting would add to the empirical discounting of the respondents. In principle, in the Multiple Profiles QALY Model no computations are required. Hence, conventional discounting is not applicable. From this follows that double discounting may happen only in the Single Profile QALY Model if conventional discounting is applied. Until now, little research has been directed to the response-bias, the stimulus-bias and the double discounting phenomenon caused by the complex role of "time" as a commodity and "time (of occurence)" as determinant of the value of commodities (see also this chapter: p. 165).

Respondents

The choice of which subjects (individual: patients, physicians; panels: general population, patient groups, experts) should perform the valuation task for the valuation of health states is a matter of concern. For prioritizing in health care for example, it is important to know whether the choice for a particular panel affects the (QALY) results, and thereby the health care priorities (see: Stronks *et al.*, 1997). The dominant view is that the choice about who should value the health states depends on the context of application of the results. In our view, two main contexts exist, i.e., individual decision making on the one hand, and societal decision making on the other hand. It is generally assumed that for individual choices the preferences of patients themselves are required, whereas societal choices should be based on the values obtained from a representive sample of the general population (in fact, the assumed bias of ones own health is partially responsible for not using patients as respondents in this particular context). However, we observed in the Laryngeal Study (see: Chapter 9) that at the group level, layman without any experience of this type of cancer, ultimately seem to arrive at the same preferences as the former patients. This finding may be interpreted as an affirmation of earlier publications (van Agt *et al.*, 1994; Krabbe *et al.*, 1997) that concluded that, given an equal level of information, which requires time consuming processes, no systematic difference exists between different types of respondents (see: Chapter 4). Stated in other words: here, quick is dirty. If confirmed, this implies that people are capable of imagining a broad range of different health states and arrive at almost identical appraisals (scaling), notwithstanding the fact that they have no personal experience of the described health states.

Analysis

Studies in this area usually focus on the development of health-status classification systems. As a rule, the properties and effects of the valuation methods used, the variation among the respondents, the health-state scenarios used, the stimulus and response modes, and the possible biases due to the design of the study are at best partially examined. The estimation of a value function by which the total set of health states of a classification system can be predicted is of course the ultimate goal of developing a health-status classification system, but we have to go a long way before all aspects are described and tested, as the EuroQol undertaking illustrates.

In our studies, we were not primarily interested in the *prediction* of the valuation of health states, as is commonly aimed at by the use of regression analysis. We were interested however in *describing* specific properties of the valuation methods, the effects of the stimuli (health states), the amount of variance attributed to occasions and respondents and so on (i.e., step 1 of Table 2, Chapter 2). This stage precedes prediction. Therefore, we conducted experiments from which the evoked data was analyzed with several variants of the statistical technique of analysis of variance. The approach of *experimental* design is essential to disentangle the multiple sources of variation which interact in the assignment of values to health states and to refine theories and hypotheses (see for an early enterprise on this topic: Elstein *et al.*, 1978). It allows for a deep exploration and description of the relationships. Yet the *survey* type of research, where data is standardly analyzed by multiple regression, seems best adapted to extensive rather than intensive research. For survey studies, the lack of control of the conditions, stimuli and of the intervening variables is often a severe handicap. Anderson

(1976) has stated: "Regression-correlation methodology, which can be so useful in applied prediction, is inadequate and misleading when it comes to testing theoretical models". As of this moment, several aspects are unclear, therefore, more specific empirical (experimental) studies are required to provide clarification about missing pieces in QALY modeling. In such studies, analysis should be based on non-aggregated data, whereas post-analytical transformation of the data, such as age-weighting, should be kept to a minimum.

By the use of Generalizability Theory (Cronbach *et al.*, 1963; Shavelson *et al.*, 1989; Streiner & Norman, 1995), a specific application of analysis of variance, we were able to reveal various sources of measurement error in the elicited values for health outcomes. A G-study excluded different response behavior in our rather homogeneous sample of students. After appropriate standardization of the individual valuations on the four different valuation methods, less than 1.0% of the total variance was found to be attributable to systematic different response behavior among the students. Although all the methods to some extent seem to be biased, the SG, TTO and RS methods yield health-state valuations that were satisfactorily reliable at the group level. These findings support the validity of constructing societal values for health states based on aggregated data. Hence, G-theory should be the primary approach in the field of valuing health outcomes.

The advocated multitrait-multimethod (MTMM) methodology (Froberg & Kane, 1989d; Hadorn & Hays, 1991) to examine the construct validity of valuation methods unfortunately proved to be not applicable due to the specific structure of the valuation data. This specific data structure is a direct result of the valuation task associated with the assessment of health states (stimulus-scaling task) and the fact that the stimuli do possess a "manifest" ordered structure. Hence, MTMM is not adequate for the analysis of health-state valuation data.

The individual perspective

Information offered to the respondent

In choosing how to offer information to respondents, a clear difference should be made between the application of the QALY approach at the individual level (clinical decision making) and at the societal level. At the individual level, the information is described in an almost standardized way by the use of naturalistic health scenarios, without or, more often, with disease labels. The information offered by this type of scenario, especially when used within the context of clinical decision making, is restricted to the health status of end-states (e.g., treatment outcomes). For "straightforward" diseases and competent patients this seems to be a good option. In practice most patients are not a *tabula rasa* but are reasonably informed about prognostic/prospect aspects and possible side-effects. It is doubtful whether they are capable of disregarding such information during the valuation task, if such a task is only directed at the valuation of *Q*.

Due to the specific setting of patients' decision making, multiattribute scenarios appear to contain insufficient information, either because of lack of attributes (domains), the lack of calibration (refinement of the response mode), and maybe by the lack of perspective and

prognosis. Moreover, if patients are to some extent already acquainted with the objective of decision making, the use of abstract multiattribute scenarios may be too impersonal.

If the decision problem is complicated, it may be preferable to use (lifetime) health profiles. Instead of the standard naturalistic scenario approach, describing solely a specified end-state, the advantage of assessing health profiles is the account for the health status during the complete time period. If such comprehensive naturalistic scenarios are used that describe the complete course of treatment (see: Chapter 8), an important aspect is the type and the amount of (clinical) information offered to the patient. The use of this type of scenario requires patients to remember and utilize ("process") many data about the various treatment outcomes, as well as diagnostic and treatment related information. In order to arrive at a consistent, well-considered valuation, one should be aware that patients may focus on some vivid, and in some cases, minor or irrelevant effect(s), instead of the more lasting impact of the treatment.

The view that patients are the best informed in regard to decision making, either individually or as societal representatives, appears not to be necessarily true. First, in the ex ante situation, patients and non-patients alike are confronted with uncertainty. Moreover, it seems that because of for example emotional constraints, some period of time (see: Chapter 9) elapses before a patient arrives at his/her "true preference". Second, it is only in rare situations that a patient is informed by own experience about *any* of his/her treatment options (e.g., try out of drug therapy). In our example, patients facing the choice between surgical treatment and radiation treatment had not experienced either of the two treatment modalities. The use of former patients for consultation (if not too selective) certainly may be helpful for those patients who have yet to decide. The argument that the patients own preference system should be the primary yardstick may still be valid but with certain caveats, i.e., under pressure and in the absence of experience, patients may be unaware of their capacity for endurance or ability to adapt to life-endangering situations. Therefore, they may be ill-represented by their own ex ante expressed preferences.

Instead of using a decomposed QALY model in clinical decision making, the Single Profile QALY Model (see: Chapter 8) could be considered as an alternative. It should be recognized however that this model, which is based on a complex and implicit cognitive weighting of a complete lifetime health profile, demands a considerable cognitive effort from patients. This effort may not be manageable under stressful conditions. In our opinion, the analytical approach (decision tree analysis) and the Single Profile QALY Model approach represent the two extremes within the area of the appraisal of treatments at the individual level. The development of alternative models and valuation tasks in between these two approaches is the current challenge.

Responses

One of the main characteristics of individual decision making as opposed to societal decision making, is that probabilities play a key role in the ex ante individual approach. Next to life years (L) and health status (Q), it is particularly the probability/uncertainty factor that is important in medical decision making at the individual level (ex ante).

Note, that for societal decision making the role of probability is exchanged for the role of number of patients affected. It is no longer individual uncertainty considerations that are relevant but distribution considerations (see: Appendix A).

With respect to the use of the QALY approach at the individual level, we are confronted with the following limitations. The first one is a formal one and deals with the question of whether the observed violations of EUT can justify the QALY approach at the individual patient level. Some empirical studies have shown violations of the assumptions made by EUT. Second, the use of SG as the prefered method of eliciting values under uncertainty is questioned. Prognosis, expressed as probabilities in the decision tree, seems to be involved twice: once during the valuation of these elements of the valuation task performed by the SG method *(p)*, and subsequently by the computation of the decision tree analysis with the probabilities of the different occurrences. "Double counting" is not an issue here (see the analogue analysis of double discounting: p. 173), but if in SG tasks themselves, the assumptions for p do not hold, it is likely that they are also violated for simple aggregation of the profile scores by p-weighting (see: Appendix A).

Introduced as an adequate strategy to incorporate individual appraisals of patients and to overcome cognitive overload during the valuation stage, the analytical approach based on decision (tree) analysis has received much interest. However, our studies demonstrated that this decomposed approach (Partially Decomposed QALY Model) may be reasonably valid at the group level, but also that on the individual level results are not reliable (see also: Chapter 9). The replication of the rather complex valuation task and rehearsing of the offered information seems to be a necessity for providing reliable preferences.

Analysis

Another limitation of the clinical decision approach at the individual level is the amount of intra-individual variation in the valuation of health states (in our laryngeal cancer study: end-states) (see: Chapters 7 and 9). Atypical valuations among patients cannot simply be handled as population variation, as is the case at the other application levels, but are regarded as distinctive, subjective preferences. In Chapter 4, the reliability of valuations for EuroQol health states elicited by the SG method, was not high (0.56). Only after averaging individual valuations (single measurements), can very reliable values for health states be attained for applications at the societal level and for Medical Technology Assessement studies. Such a strategy is not possible at the individual level, where the information of interest is the individual "real score" itself. Individual related biases cannot be avoided by a more appropriate design nor adjusted by analysis. In the case of single measurement, even no mean and standard deviation is available to estimate a reliability coefficient. This is one of the reasons, why researchers and physicians are cautious about the practical implementation of clinical decision analysis (Roest *et al.*, 1997). Sensitivity analysis is the only feasible and valid tool to explore the effect of imprecise individual valuations.

Future perspective

The limitation of disease-specific non-standardized quantification of health outcomes have inspired many health services researchers to develop strategies to quantify this value in a

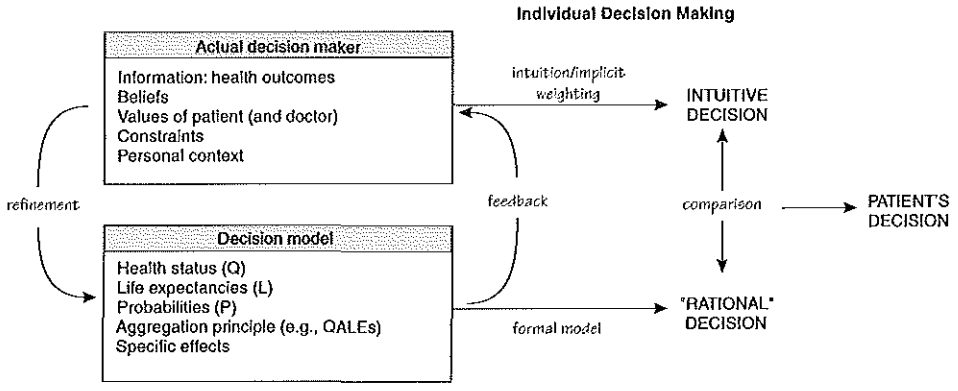


Figure 2
A schematic representation of the position of clinical decision analysis as part of a counseling process

standardized way. This unifying trend can also be observed in fields like educational achievements and environmental monitoring. The dominant approach in health care to deal with this issue is the QALY model.

As the general QALY model appears to be a useful tool and as the concept on which it is based allows for further improvement and testing, we expect continuous dissemination of this tool. Hence, the state of the art of the QALY approach today is probably best viewed as a preview on (a more) rational choice process at the individual and societal level.

Individual perspective

So far, the formal decision analysis approach can only be part of a clinical counseling procedure (Figure 2). It is a supportive rather than a prescriptive technique to evoke the ultimate decision (Nease, 1997). The apparent difficulty of assigning numerical values, which hold individual validity, to specific health conditions is a major limitation of clinical decision analysis. We may add that the limited experience of patients to process the different elements of information (health status, life expectancies, probabilities of success) and the stressful circumstances under which such valuation takes place, to a large degree prevent the use of any formalized decision approach at the individual level. It is not the validity of the clinical decision paradigm itself, but the reliability of the individual preferences that forms the weakness of the approach.

New developments based on the incorporation of the complete valuation task into multi-media applications together with the opportunity to link to central data-bases and the use of iterative strategies to reach optimal preferences, may in the end contribute in a practical way to decision making at the individual level.

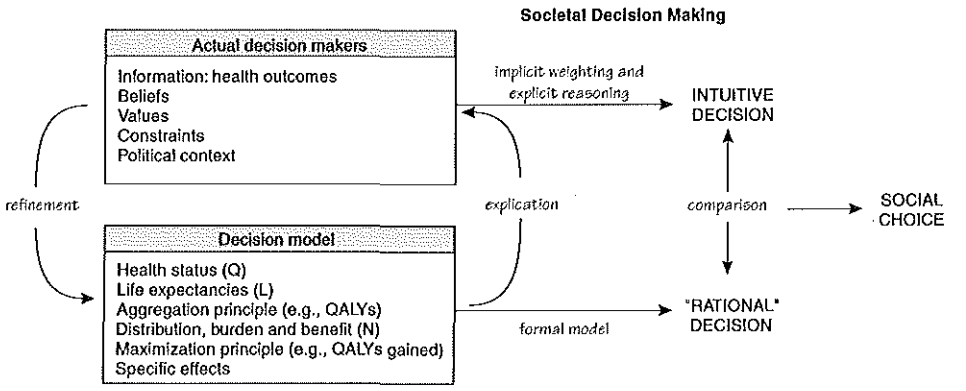


Figure 3
A schematic representation of the position of societal decision making

Societal perspective

Monitoring the health conditions of a populations will probably receive more attention during the coming years. Not only governments encourage comparison within countries, but larger entities, such as the European Community will want to monitor the health of larger populations and compare between countries. In fact this trend is already visible. Recent pioneering projects carried out by inter-governmental organizations such as the World Health Organization and the World Bank will be followed up by large scale studies in the United States and the European Community.

In the near future, registries will make available more data relating to incidence and prevalence of diseases together with information on associated life expectancy and other health care information. For many diseases it will then be possible to improve the precision of quantifying their impact by using approaches such as the QALY (see also: Figure 3). The combination of cost-utility ratios (QALY league tables) and increased knowledge about the public's perception of health may lead to better health policy decisions.

Conclusions

- Due to the specific task of valuing health outcomes, e.g., stimulus-scaling task, some particular psychometric techniques (i.e., multitrait-multimethod approach, Cronbach's alpha) are not applicable, as they are intended for standard response-scaling tasks.
- Little is known about the specific effect of statements about duration in relation to the valuation of health outcomes.
- Compared to tradeoff valuation methods, assessments of health states elicited by the rating scale method proved to be more reliable. Their validity, presumably, is lower.
- More studies are needed on multiattribute health-status classification systems and their associated methodology in order to arrive at value functions.

- The validity of the QALY approach primarily depends on reliable health-outcome values and less on its conceptual properties.
- Experimental studies are needed to define empirically some of the required standards in the field of health-outcome valuation.

Der ist der glücklichste Mensch,
der das Ende seines Lebens mit dem Anfang in Verbindung setzen kann.

GOETHE
(1749 – 1832)

Appendix A

"... de diviser chacune des difficultés que j'examinerais, en autant de parcelles qu'il se pourrait, et qu'il serait requis pour les mieux résoudre"[*]

R. Descartes

QALY Models: a Formal Approach

Introduction

The assignment of a utility or value to health outcomes, or, technically defined, to life prospects, is a difficult cognitive task. In order to deal with the valuation of such life prospects, procedures have been developed that decompose the prospects into elements prior to empirical valuation. This may be interpreted as a simplification of the "subject" to be valued. After each element of the prospect is separately valued, these values are subsequently combined baccording to certain rules to arrive again at the value of the life prospect as a whole. The above description of the decomposition of health outcomes rests on a Cartesian assumption of "divisibility" of life prospect.

"Life prospect" may also be referred to as "prognosis", expected "health outcome" or "life-profile", though the latter two terms are less clearly defined and lack the clear *ex ante* connotation of pro-spect and pro-gnosis. "Life prospect" represents our concept of health in the context of decision making and evaluation. This concept is deliberately focused on future outcomes, rather than on past or actual health status.[†]

In this appendix, we present a generalized mathematical model for health outcomes, which enables us to derive the four QALY models, as presented and discussed in Chapter 2. We start with the introduction of a concept, the QALY-ple, which we denote as the basic unit of the QALY model. Based on this basic unit we give an example to clarify the one-to-one relationship between a formalized description of a particular health outcome and its counterpart: the values of the distinct health elements of the initial health-outcome description. Thereafter, all four QALY models will be discussed in a mathematical manner. Finally, two specific aspects will be discussed that were not raised in Chapter 2. First, we formalize the difference between the Multiple Profile QALY Model from the societal

---

[*] ("... to divide each problem that I investigate into as many parts as possible and are necessary to arrive at their best solution."), René Decartes (1596-1650) describing the second of his three methodological rules in *Discours de la Méthode*, first published in 1637 at Leiden, the Netherlands.

[†] A remote consequence of this concept is the disclaimer of past health history or actual health status on values assigned to health gains in future, which has been proposed by those aiming at Rawlsian justice (Rawls, 1997).

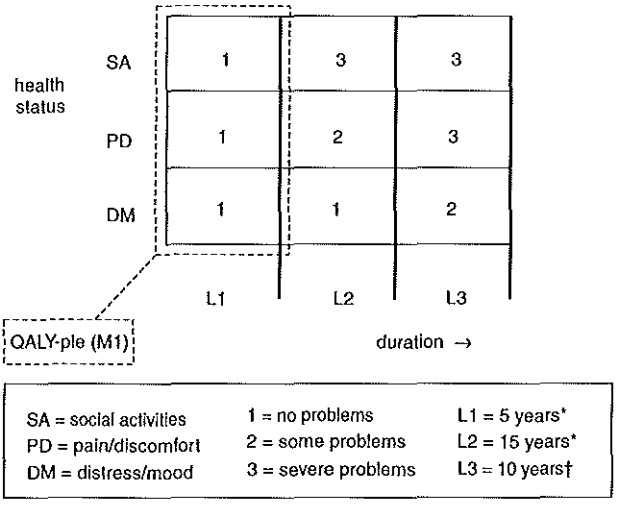


Figure 1
Representation of the description of three QALY-ples, together representing a lifetime health profile

perspective compared to the same model from an individual perspective. Second, we present a brief overview about strategies to adjust for effects due to prognostic elements not incorporated in the four basic QALY models.

The QALY-ple

We will hereafter refer to the *QALY-ple* as the smallest unit resulting from the decomposition of health outcomes. A *QALY-ple* is a discrete period, uniquely defined by:

i) constant health status (*Q*), optionally subdivided into a set of attributes (*A*) cf. some classification;
ii) duration (*L*) expressed in numerical, semi-quantitative or qualitative terms;
iii) four possibly censoring endpoints *(C)*, respectively: dead (0), alive (1), unknown or uncertain (2), unknown or uncertain although options defined (3).‡

As stated, the QALY-ple is assumed to be the smallest unit of health outcome. Starting from this unit we can build up combinations of QALY-ples ending with life prospects, in their unspecified form referred to as *health profiles*. Let *M* uniquely defined *QALY-ples* make up for a *QALY-ple chain*, where c_1 through $c_{M-1} = 1$. Furthermore, a chain represents an individual (remaining) life, if for the last QALY-ple holds $c_M = 0$. QALY-ple chains will be referred to as

‡ The orientation of the future prevents the inclusion of a similar indicator to cover past or actual health.

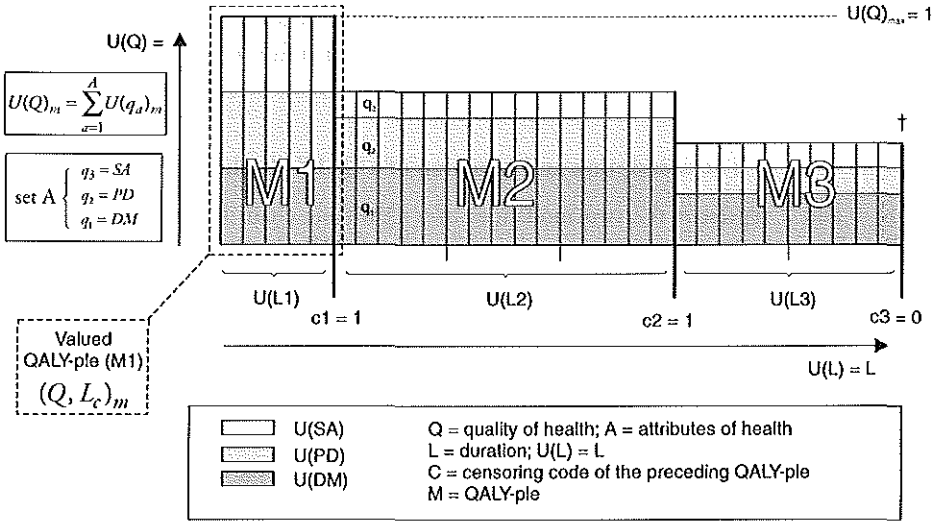


Figure 2
Representation of the valuation (the graded surface represents the value of the health outcome expressed in QALYs) of the QALY-ple chain for the Fully Decomposed QALY Model (model I)

"episodic health profiles" if $c_M \neq 0$. If $c_M = 0$ we will refer to *"lifetime health profiles"*, following Gafni (1995). Torrance & Feeny (1989) refers to "life profiles".

Lifetime health profiles can be further combined in two different ways, depending on the perspective taken. In both cases we refer (technically) to "QALY-ple trees" and more generally about "life prospect".

From the *societal* perspective, we can combine *N* lifetime health profiles to a QALY-ple tree, as a representation of the possible or realized life prospects of N individuals sharing some common health characteristics and/or some common starting point in time. From the societal perspective, it is generally not relevant whether we look ex ante or ex post, because probability as perceives by the individual plays no role. Such combined lifetime health profiles may be used in comparative studies in public health (Murray & Lopez, 1996) or in clinical intervention studies (Bonsel *et al.*, 1992); in both cases the total number of QALYs (to be defined more precisely later) represents the value of such combined lifetime health profiles.§

From the *individual* perspective, the same *N* lifetime health profiles of a relevant reference group of N patients may be similarly combined into a QALY-ple tree as a representation of a – necessarily – ex ante life prospect, from some common starting point in time, where each health profile has probability $p = 1/N$.‖

§ Without additional assumptions, it is generally not true that the sum of *N* individually valued life prospects/ex ante prospects is equivalent to the directly assigned (ex post or ex ante) value of *N* lifetime health profiles (this refers to a second divisibility assumption in utility measurement).

‖ As with the societal perspective, it is generally not true that the sum of *N* life prospects, weighted by their

In the following section we will graphically explain the differences between a description of a QALY-ple chain and its value, using a simplified three attribute, three level, health-status classification.

The description and valuation of a QALY-ple chain

Figure 1 and 2 represent the two modalities of a QALY-ple chain (here a lifetime health profile). Figure 1 merely represents a *description* of a QALY-ple chain consisting of three QALY-ples, i.e., M1, M2 and M3. No reference to any valuation is made. The duration statements (L1, L2 , L3) can be expressed quantitatively, as in this conventional example, but can also be expressed qualitatively (see later). Health-status attributes are expressed as aspects of Q, where $q_1 \ldots q_A$ are the attributes to be specified and where A is the number of attributes defined. The health-status attributes may, of course, be subdivided into more than the three levels of our example.

Figure 2 is a graphical representation of the *value* of this QALY-ple chain conform, defined here as, the Fully Decomposed QALY Model (see: Chapter 2, p. 24). The graded surface ("area under the curve") represents the undiscounted utility expressed in QALYs as unit of measurement. This figure represents the most disaggregated QALY model. As described below, we can decide the extent to which we divide a life prospect into separate elements. We propose the following notation to facilitate the discussion on the assumptions and relations between elements, dependent on the level of division.

Let a QALY-ple be described by $(Q, L_c)_m$, with its associated utility $U(Q, L)$, then the individual lifetime profile, assuming a *certain* prospect, can be described as follows:

$$\left[(Q, L_{c\neq 0})_{m=1}, (Q, L_{c\neq 0})_{m=2}, \ldots, (Q, L_{c\neq 0})_{m=M-1}, (Q, L_{c=0})_{m=M}\right] \qquad (1)$$

with its associated utility $U(Q, L_c)_M$, further denoted as $U(\mathbf{Q}, L)$, see also Figure 2.

In the preceding text, lifetime health profile and life prospect refer to "real life". In the valuation process that is described in this thesis they are the *stimulus* (stylized in some way) to be valued.

Note, that in many valuations methods, the *response* modality also contains some (artificial) health description as a comparator, or as part of a so-called calibrator. If the response modality is a rating scale, the anchor points (the comparators) are timeless, e.g., "best imaginable health state". If the response is in the form of a tradeoff task, the response modality is structured as an episodic health profile, a lifetime health profile (both time-tradeoffs) or a life prospect (standard-gamble). As the trading implies giving up some positive aspect of this health profile (acceptation of a reduced duration, or some mortality risk) until indifference between value of the "real" stimulus and the "artificial" calibrating health profile (the

probability $1/N$, equals the value assigned directly to the QALY-ple tree as a whole.

response) is reached, we refer to a health profile as a calibrator, rather than a comparator or anchor. It is important to recognize that stimulus and response are entirely different entities.

Four QALY models with different aggregation levels

We will present four possible life prospect decomposition (or disaggregation) levels in ascending order of aggregation, starting with the most commonly used model. These decomposition levels will be indicated as QALY model I to IV. Each level of aggregation is associated with specific assumptions, a formula, and an associated aggregation rule to arrive at a summary measure for health.

Each step of disaggregation assumes the conceptual validity of further division into health elements and time periods. Each step of disaggregation also implies that the resulting elements are valued separately, and that the value of the combined elements in some way has to be determined. The valuation method may differ according to the elements included, as demonstrated from literature evidence.

For such models it generally holds that the less aggregated they are, the more valid the values from the empirical task.[¶] The key concept of any disaggregated QALY model is: 1) that useful and valid valuation methods exist, suitable for the valuation of the elements and suitable for the anticipated aggregation by rules (this choice will be discussed later), and 2) that such *rules* or, more technically (based on estimation model), such aggregation *functions* exist. These aggregation functions, can be simple ("additive function") or complex. They may be tested against competing functions in some empirical way. The superiority of such a function can be conventionally established by examining predictive power, robustness, degrees of freedom left, and, to some degree, theoretical attractiveness.

Frequently however, the aggregation function appears to be considered as prior assumption, without empirical testing of its assumed qualities. If this is the case, such "function" with its associated procedural values and weights is normative, say, political, with consequent limited scientific opportunities to test its validity.

Model I - Fully Decomposed QALY Model

In the Fully Decomposed QALY Model (see: Chapter 2, p. 24), the following assumptions must hold (indices kept to a minimum; *f, g, h* and *i* representing the combination rules),

$$U(Q,\ L) = f[U(Q,\ L_{c=1})_{m=1}, U(Q,\ L_{c=1})_{m=2}, \ldots, U(Q, L_{c=0})_{m=M}]. \qquad (2)$$

Expression (2) states that a health profile consisting of multiple QALY-ples can be aggregated by the application of a function *(f)* on each of the separately valued QALY-ples. The

¶ We are aware that even the model without aggregation should be subjected to the question of validity: does the *hypothetical response behavior* really represent preferences of desirability, in a situation where neither a patient nor society is able to experiment in real life with the options available?

conventional QALY assumption is that an *additive* function aggregates the values of the separate QALY-ples.

In addition, expression (2) may include a constant proportional risk posture assumption. It *may* also include a discounting effect, if deemed necessary. (Both technically simple extensions of the general QALY model are not further discussed in this appendix).

$$U(Q, L_c)_m = g[U(Q)_m, U(L_c)_m] \tag{3}$$

Expression (3) states that the two distinct elements (Q and L) of a QALY-ple can be integrated by a function *(g)*, where g is usually defined as a simple *multiplicative* function. With the following two formulas, we further expand on *U(Q)* and *U(L)* respectively as follows:

$$U(Q)_m = h[U(q_1)_m, U(q_2)_m, \ldots, U(q_A)_m]. \tag{4}$$

Expression (4) states that the separate attributes ($q_a \ldots q_A$) of a health-status concept (Q) can be integrated by a function *(h)*, where h is defined as either an *additive* or a *multiplicative* function with $0 \leq h \geq 1$. We assume h to be additive, unless stated otherwise.

Function h can be specified after the separate assessment of the attributes. Another frequently used approach consists of using an existing multiattribute classification system. Such a classification system offers a value function (h) based on the results of a prior study dealing with the estimation of the contribution of each attribute (q_1, q_2, ..., $q_A = Q$). The estimation of such value functions is discussed in Appendix C.

Expression (5) states that the appraisal of L (survival, life years) can be different from the actual number of years. The function i can be different for each separate QALY-ple and can also be altered due to the censoring end-point c.

$$U(L_c)_m = i(L_c)_m \tag{5}$$

We further assume $i(L_c)_m = L$, irrespective of c and m.

Note that a time index (as a preparation for discounting) is omitted at this stage. Note also that in empirical studies, such as the ones described in this thesis, commonly q's or Q are valued under the assumption that the c of $L_c = 2$.

In general, the Fully Decomposed QALY Model defines *U(Q, L)* as follows (optional constant terms are left out of all the formulas presented):

$$U(\mathbf{Q}, L) = f(g(h(Q)_m, i(L_c)_m)) = \sum_{m=1}^{M}\left[L_{c,m} \times \left(\sum_{a=1}^{A} U(q_a)_m \right) \right]. \tag{6}$$

Clearly $\sum_{m=1}^{M}$ and $\sum_{a=1}^{A}$ represent additive combination rules. Note, that if we use a multiplicative rule for *h*, the formula changes into:

$$U(\mathbf{Q},L)=\sum_{m=1}^{M}\left[L_{c,m}\times\left(\prod_{a=1}^{A}U(q_a)_m\right)\right]. \qquad (7)$$

If the functional form of *f, g, h, i* rests on empirical evidence, the model is *descriptive*; however we may wish to define one or more of these functions as an assumption making the model (partially) *normative*. The "best" empirical function often is difficult to define. For example, the EuroQol Group reported that when considering function *h* at the group level the multiplicative model did not fit better than the additive model (van Busschbach *et al.*, 1997).

A common error in argument arises when the particular choice of a valuation method (TTO, SG, WTP, VAS, PTO) for *h* is based on conceptual considerations relevant to function *f* or – see below – function *k* and *l*, rather than considerations relevant to *h* itself (we have called this the template error; see: Chapter 10, p. 173).

The most frequently applied functions for *i, f, g, h, k* and *l* were described in the preceding part. With the phrasing "most frequently", however, we are not stating that those functions are the best. It should be kept in mind that, except for function *h*, very few studies have been conducted to investigate the empirical relationships and properties of these functions.

Model II - Partially Decomposed QALY Models

This model (see: Chapter 2, p. 25) differs from model I, in that assumption (4) is no longer deemed to be necessary. Hence, further divisions of *Q* into attributes which are to be separately valued, is not required. With this model, we no longer depend on analytically described *Q* by means of some multiattribute health-status classification system (like EuroQol), but we can also apply so-called "naturalistic" descriptions or simple disease labels like "migraine" as in Murray's approach. The general formula for this model is:

$$U(\mathbf{Q},\ L)=f(g(Q,L))=\sum_{m=1}^{M}L_{c,m}\times\left(U(Q)_m\right). \qquad (8)$$

If we look for valid descriptive functions for all possible health states, this model demands that all possible descriptive labels have to be valued. If we don't assume that $U(L)_m = L$, this model changes into:

$$U(\mathbf{Q},\ L)=f(g(Q,L))=\sum_{m=1}^{M}[U(Q)_m\times U(L_c)_m]. \qquad (9)$$

If (3) and (5) are no longer assumed, the empirical task implies the assessment of $Q_m L_m$ (QALY-ple) rather than Q_m, with consequent reduction of the aggregation procedures. Figure 3 provides a description of this model, which is represented in the following formula:

$$U(\boldsymbol{Q}, \boldsymbol{L}) = f(Q,L) = \sum_{m=1}^{M} U(Q, L_c)_m. \tag{10}$$

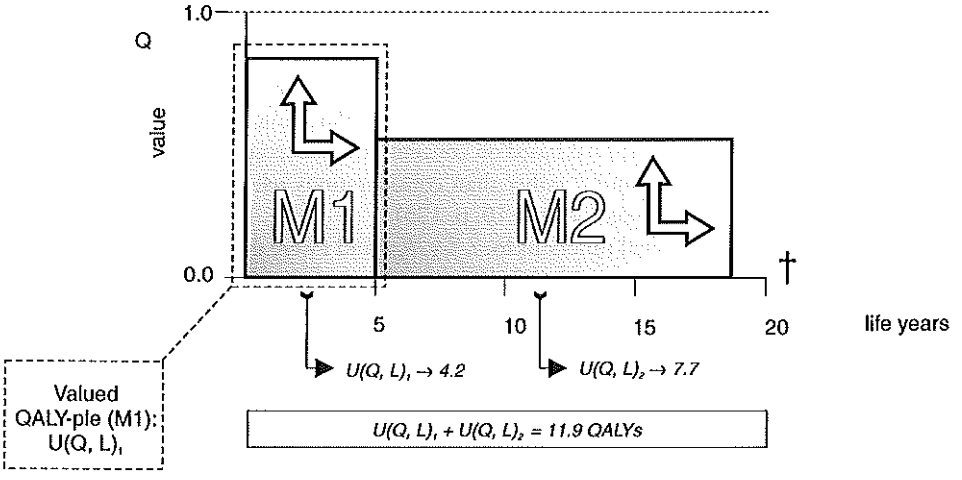


Figure 3
MODEL II — Conceptualization of the Partially Decomposed QALY Model.
Note that this graphical presentation gives only an impression of the weighting strategies of respectively the elements Q and L

Data obtained and analysed using this model allows us to test the assumption $U(L)_m = L$ for all m. Moreover, this model or more aggregated models are required in cases of non-numerical descriptions of L.

Model III - Single Profile QALY Model

Finally we may omit assumption (1), which leaves us with a full health profile to be valued:

$$U(\boldsymbol{Q}, \boldsymbol{L}) = U\begin{bmatrix}(Q, L_{c\neq 0})_{m=1}, (Q, L_{c\neq 0})_{m=2}, \ldots, \\ (Q, L_{c\neq 0})_{m=M-1}, (Q, L_{c=0})_{m=M}\end{bmatrix}. \tag{11}$$

Single integral profile modeling (see: Chapter 2, p. 28) has been advocated among others by Mehrez and Gafni (1989; 1996), when they introduced the Health-Years Equivalent concept.[¥] Their concept must be regarded as a QALY Model III concept.[++] Despite the inherent difficulties of this cognitive task, the attractiveness of this model is the use of more "realistic" stimuli to be valued, assuming less distortion from the hypothetical test situation in general. Data obtained according to this model allows for testing the assumptions of Model II, expression (10), that $U(Q, L_c)_m$ is independent from c and from m.

Model IV - Multiple Profile QALY Model: life prospects from the societal and individual perspective

So far, the value or utility of individual periods of life ("episodic/lifetime health profiles") has been modeled in instances where all time periods were assumed to be certain. Now we will focus on multiple profiles (see: Chapter 2, p. 30), assuming that some reasonable common starting point is available. In this case the array of outcomes can be viewed from two perspectives, as explained before.

Societal

The societal perspective views multiple health profiles as separate life histories and we can imagine the complex task of valuing an array of N outcomes of N individuals. A societal model for N outcomes is:

$$U^{societal}(Q, L)^{N^*} = k[U(Q, L)^1, U(Q, L)^2, \ldots, U(Q, L)^N] \qquad (12)$$

with k defined as some *general* function. Conventionally k is *assumed* to be a simple additive function κ ("one man, one vote", without reference to whom benefits occur):

$$\kappa[U(Q, L)^1, \ldots, U(Q, L)^N] = \sum_{n=1}^{N} U(Q, L)^n \qquad (13)$$

Equivalence methods, such as the person-tradeoff (PTO) method (Nord, 1995; Murray & Lopez, 1996) are, in principle, suitable for the task of determining k empirically, hence checking the validity of κ. We may also define k normatively for the inclusion of distributional characteristics.

Distributional optimality usually is defined as the highest degree of homogeneity in either health gain and/or the degree of actual or instantaneous health per se. If operationalized, we

¥ "Healthy-Years Equivalents" are not unlike "QALYs", "DALYs" or "function-years", and are not connected to a particular method of valuation or a specific level of aggregation.

++ Gafni and Mehrez used the TTO method for valuation of such profiles, but other tradeoff and non-tradeoff techniques could have been used.

may arrive as a measure of outcome, corrected for inequity, at, for example, $\kappa[U(Q, L)^1, U(Q, L)^2,...,U(Q, L)^N] = N \times min\{U(Q, L)^1, U(Q, L)^2,...,U(Q, L)^N\}$ to reflect a societal viewpoint which is based on the least outcome gain. If we wished to adopt a Rawlsian approach*, we may define $\kappa[(U(Q, L)^1, U(Q, L)^2,...,U(Q, L)^N]$ as $N \times [U(Q, L)^{\Delta}]$, where delta indicates the chain belonging to the person with, doing nothing, the largest loss of QALYs compared to some absolute reference. Following others, delta should point to the person with the lowest Q at the time of the decision. Even functions of the sort $\kappa[U(Q, L)^1, U(Q, L)^2,...,U(Q, L)^N] = j(N)\sum_{n=1}^{N} U(Q, L)^n$ are possible to account for the number of persons/patients involved. Given specification of k, parametrization can be achieved by PTO or other equivalence techniques conducted by an appropriate design.

Individual

The individual viewpoint should be distinguished from the societal viewpoint, as follows. Personal *uncertainty* about which lifetime health profile the individual will experience, is (conventionally) quantitatively reflected by the probabilities in the decision tree. The individual utility model for an array of N possible lifetime health profiles can therefore be represented as follows:

$$U^{individual}(Q, L)^{N^\circ} = l[U(Q, L)^1, U(Q, L)^2,...,U(Q, L)^N]. \qquad (14)$$

Here, $(Q, L)^{N^\circ}$ indicates N lifetime health profiles, l indicates some general function.

In the conventional individual expected utility model, we assume for $l(U)$ a simple additive weighting function similar to κ:

$$\lambda = \frac{1}{N}\sum_{n=1}^{N} U(Q, L)^n. \qquad (15)$$

If we chose λ for l, with the Fully Decomposed QALY Model as described in Chapter 2 and (6), we arrive at:

* The philosopher Rawls (1997) has introduced (first published in 1971) a conception of justice (i.e., equity), within a political context, using the "veil of ignorance" thought-experiment that leads to an assumed universal conception of justice. The rationale of this idea is that any prior information (e.g., own place in society, class position, natural assests and abilities, health, prognosis) that may affect a person's choice in favour of themselves or particular parties should be ignored.

$$U^{individual}(Q,\ L)^{N^\circ} = I\Big(f(g(h(Q^A)_m, i(L_c)_m))\Big) = \sum_{n=1}^{N} \frac{1}{N} \left\{ \sum_{m=1}^{M} \left[L_m \left(\sum_{a=1}^{A} U(q_a)_m \right) \right] \right\}^n \qquad (16)$$

However, a more subtle role of probability and prospect (prognosis) may be assumed.

Extension of the QALY model to account for prognostic considerations

Prognosis in this formal context has two components:

- A probability of death in some (personal/individual) evaluated future; the degree of awareness or certainty about its occurrence apart from health status considerations.
- The health-status prospect, that is the individually foreseen future development of health status per se after the defined QALY-ple, or some other personal time horizon, apart from survival considerations.

We may deal with the latter problem by ignoring "prognosis", conventionally weighting each observed lifetime profile by *1/N* only, see (15) and (16). Otherwise we may incorporate it as follows

$$\underbrace{U(L_{c,m}) = \gamma[U(L), U(C)]}_{\text{instant time utility modifier}}\ (I) \quad \text{or} \quad \underbrace{U(Q_{c,m}) = \delta[U(Q), U(C)]}_{\text{health status utility modifier}}\ (II)$$

or incorporate its effect into the already mentioned function, λ (III).

For each strategy we will give an example. We are nevertheless aware that more complex functions are required to deal with interaction effects between adjacent QALY-ples.

I $$\gamma\,[U(L), U(C)] = L^{\alpha}\ ,\ \alpha \text{ independent of } C. \qquad (17)$$

II $$\delta[U(Q), U(C)] = \left[\sum_{a=1}^{A} U(q_a) \right]^{\alpha_a}\ ,\ \alpha_a \text{ independent of } C. \qquad (18)$$

An example of prognostic adaptation (risk aversion) incorporated in λ itself is as follows:

III $$\lambda[U(Q,L)^1, \ldots, U(Q,L)^N] = min\ \{U(Q,\ L)^1, \ldots, U(Q,\ L)^N\}. \qquad (19)$$

Note that the distribution effects in the societal model emerge as the prognosis effects on utility in the individual utility model.

Valuation methods and aggregation functions

Various valuation methods or, technically, stimulus-response methods (standard-gamble, time-tradeoff, visual analogue scale, equivalence methods, scaling approaches) exist which are able to assign a value to a multiattribute stimulus. "Multiattribute" can be interpreted as the attributes combined in the functions *i, f, g, h, k* and *l*. For *i* the attributes are time periods (usually defined in a quantitative way), for *h*, the attributes are the number of health-status dimensions, for *g* the combination of health status with duration, etc.

Strictly speaking, the nature of a particular valuation method does not matter provided that:

- The valuation method is operationalized so that it fits the attributes of the stimulus.
- The exact definition of the comparator requires more attention the more the function to be studied is aggregated to avoid inconsistency of dimensionality of stimulus and response.
- The method is sensitive to variations of these attributes.
- The response evoked by the valuation methods possess interpretations with a clear "health value" connotation.
- The response, if necessary, is appropriate for subsequent processing in the QALY model chosen.

The values obtained by the use of the valuation methods can be analyzed, using (from a statistical point of view) the function that most efficiently explains the effect of the attributes on the summary value.

Appendix B

Systeme ohne Facta sind leer,
Facta ohne Systemgedanken sind blind.
K. Bühler

Human Information Processing

Anderson's theory (1971) provides a general approach to the analysis of the contributing effects to any experimental stimulus-response process, here, health scenarios and their elicited values. If we think of the observed stimulus information (X_i and Y_j) as particular levels i and j of two attributes X and Y of a health state, the cognitive process involved in its valuation may be thought of as follows (see: Figure). First, the respondent transforms each piece of information into a subjective stimulus value (S_Y, S_X) by process step I. In the next step the respondent applies a *cognitive* rule (II) to transform these scale values into a subjective response, Y_{YX}. The last step (III) involves the respondent transforming this subjective response into an observed response, R_{XY} (Veit *et al.*, 1982). There is a close resemblance between this multiattribute theory of human information processing and the fundamental relationship that Stevens (1957) developed in the field of psychophysics. According to Stevens, the fundamental psychophysical relationship (cognitive rule) is:

$$\psi = f^*(S), \tag{1}$$

where S is the physically measured intensity of a stimulus, ψ is the resulting psychological magnitude, and f^* is the "psychophysical" function that transforms one into the other.

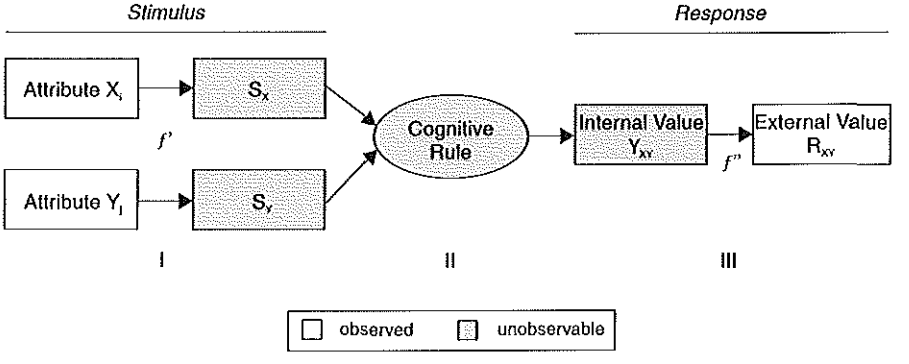


Figure
Human information process for the valuation of a two-attribute health scenario

A similar function is assumed for the response, R, by means of a second "psychophysical" transformation

$$R = f^{**}(\psi). \tag{2}$$

Because the intervening variable, ψ, is not itself observable, the responses of the subjects can at most determine the form of the single, overall "physical-physical" relationship

$$R = f^{***}(S) = f^{*} \cdot [f^{**}(S)], \tag{3}$$

in which ψ does not explicitly appear (see also: Shepard, 1981). If the appraisal of health states is based on comparable mechanisms there is no reason to assume in advance that such responses (values) have underlying interval properties (e.g., linear response function). Only complex experiments with carefully designed (factorial) designs of the stimuli and other co-factors in such cases are capable of disentangling the composed response (R) into the two distinct underlying effects, f^{*} and f^{**}.

Appendix C

No object can be defined as timeless and, consequently, the study of time should be part of human experience and behavior.
P.A. Vroon

Mathematical Value Functions for Multiattribute Health States

Introduction

A basic assumption of the Fully Decomposed QALY Model is that the separately valued attributes $(q_1, q_2, \ldots, q_A)$ of the applied health concept add up to 1.0 for Q as a whole if all health-status attributes are optimal. Different functions, called value functions, are available to aggregate the separate values of the attributes to a summary value for Q. For the two models based on a higher degree of aggregation (e.g., the Partially Decomposed QALY Model and the Multiple Profiles QALY Model), such a specific function for Q is not required, for the reason that Q is valued integrally with other health outcome elements (see: Appendix A).

Thus, the use of the Fully Decomposed QALY Model, requires the postulation of a specific estimation function specifying the relationship between the values of the attributes (denoted as function h in Appendix A). This relationship can be described by a simple additive main-effect function (without interaction terms):

$$U(Q) = c + w_1q_1 + w_2q_2 + \ldots + w_Aq_A. \quad (1)$$

Another possible function is the multiplicative main-effect function:

$$U(Q) = c \times w_1q_1 \times w_2q_2 \times \ldots \times w_Aq_A. \quad (2)$$

Model (2) is less popular, mainly because it is more demanding from a computational point of view, given the same response data.

So far, experience with models based on a transformed response value is limited, despite their wide use in other scientific areas (e.g., Y = logit (Q) or Y = log (Q)). The logit transformation is the standard model for event-modeling in epidemiology and is convenient for response data with a bounded (0 – 1) distribution. The log transformation provides a model which is an alternative for the model described in (2). We refrained from discussing models based on non-linear functions and other special functions (e.g., optimizing the predicted median rather than the mean).

One popular model is the Partially Decomposed QALY Model. In cases where naturalistic health-state scenarios are used, the only requirement is the separate valuation of all relevant scenarios. The other is the Fully Decomposed QALY Model. The attractiveness of this latter model is that it can be used together with one of the available multiattribute health-status classification systems. The separate *valuation* of the health-status attributes (q_a), here the estimation of the set of attributes $(q_1, q_2, \ldots, q_A)$, is conducted in a prior study, specifically carried out to estimate the relative contributions of the separate attributes and their levels. This is known as: the estimation of the *value function*. If an appropriately valued function for a classification system exists, such a system allows the obtaining of values for health states in a straightforward manner. In order to obtain values for a particular health state, the only requirement is to classify the state according to the system and subsequently to apply the value function. Several approaches are available to conduct studies to estimate value functions (see: Chapter 2, p. 40). Ideally, the value function to arrive at a single value for a health state (under the assumption of independence from duration) should represent the general cognitive rule that is used by human beings in processing and assessing the multiple attributes. In this context the cognitive rule should reflect "overall ex ante preference", the assumed heuristic used by individuals in comparative health decision. This heuristic is not necessarily the same as a health *care* decision, which may be more validly approached by ex post judgments and which may account for more aspects than health [change] alone. A general model for the process of preference measurement is presented in Appendix B. As a ratio measurement level is assumed, the value of the difference between two health outcomes is assumed to be reflected by the difference of the two preference values assigned to the response outcomes.

The estimation of such a value function consists of a two-step procedure. First, a predefined set of multiattribute health-state scenarios is valued. Second, by using one of the available modeling techniques (see: Chapter 2, p. 39) a value function* is estimated.

Different functional relationships between the attributes are allowed to construct value functions, each one assuming specific functional relationships between the health-status attributes and, if required, specifying the scale *within* (levels) an attribute. Extensions and combinations of the two simple basic models underlying (1) and (2) are possible. The most extended model in this context is the multilinear (utility) function†, here presented in an example comprising only three attributes (q_1, q_2, q_3), similar to the example in Figure 1 and 2 in Appendix A:

$$U(Q) = U(q_1, q_2, q_3) = c + w_1 q_1 + w_2 q_2 + w_3 q_3 + w_{12} q_1 q_2 + w_{13} q_1 q_3 + w_{23} q_2 q_3 + w_{123} q_1 q_2 q_3. \quad (3)$$

* A value function is not identical to a prediction function. The value function predicts the full set of health states of the classification system, whereas the latter function is solely focused on the prediction of the data of the data set (e.g., the valued health states).

† Another name used to refer to this model is the fully factorial model.

In this example comprising only three attributes, the number of interaction terms is modest, three first-order interactions and one second-order interaction. If the number of attributes increases, the number of interaction terms in the (saturated) value function increases rapidly.‡ For four attributes the total number of interaction terms is 11, for five attributes it increases to 26 and for six attributes the number of interaction terms already becomes 57. As several empirical studies have shown that for most health-status value functions the contribution of second and higher-order interaction terms is negligible, restricted models are used. For instance, the additive main-effect model (1) extended with first-order interactions has been used in several studies. For the situation comprising three attributes such a model is:

$$U(Q) = U(q_1, q_2, q_3) = c + w_1 q_1 + w_2 q_2 + w_3 q_3 + w_{12} q_1 q_2 + w_{13} q_1 q_3 + w_{23} q_2 q_3. \quad (4)$$

Functions within the MAU model

Confusion has surrounded the term "multiplicative" model. Originally, Keeney and Raiffa (KR) introduced three different utility functions: I) the additive function, e.g., (1), II) the KR-multiplicative function (another one than (2), see below), and III) the genuine multilinear function (Keeney & Raiffa, 1976).

The first function is identical to the simple additive function known in behavioral and social sciences. The function has been used several times in deriving a value function for multiattribute health-status classification systems.

The second function, the KR-multiplicative function (Keeney & Raiffa, 1976: p. 238), is not similar to the basic multiplicative main-effects function (2). The confusion about the term "multiplicative" originates from the following. Initially, Keeney and Raiffa formulated a function that combined the characteristics of both the additive (main-effect) function (1) and the multilinear function. This function reads as follows:

$$U(Q) = U(q_1, q_2, q_3) = c + w_1 q_1 + w_2 q_2 + w_3 q_3 + w w_1 w_2 q_1 q_2 + w w_1 w_3 q_1 q_3 + w w_2 w_3 q_2 q_3 + w^2 w_1 w_2 w_3 q_1 q_2 q_3. \quad (5)$$

If we restrict w to $w \neq 0$, then multiplying with w, adding 1, and, finally factoring, yields the following expression:

$$U(Q) = wU(q_1, q_2, q_3) + 1 = \prod_{i=1}^{3} [w w_i(q_i) + 1]. \quad (6)$$

‡ The number of interaction terms is $2^n - n - 1$ (n = number of main effects).

This "multiplicative" function, further reduces to the additive function (1) if $w = 0$. So, it is due to the restrictions imposed on equation (5) that the equation that is obtained looks as though it is a multiplicative model. In fact it should be qualified as a quasi-multiplicative function.

The third function described by Keeney and Raiffa is a generalization of both the additive and the quasi "multiplicative" function and is described by equation (3), in addition adding a restriction (the sum of *ws* equals 1) on this equations. In cases this restriction is omitted, (not the case in the MAU models) we are dealing with the full factorial model (i.e. all possible interaction terms included).

Examples from current multiattribute health-status classification systems

In this section, we briefly discuss the three main multiattribute health-status classification systems currently available. These are: the Canadian Health-Status Classification system (HSCS), later renamed as the Health Utility Index Mark-I and its successors, the Health Utility Index Mark-II and Mark-III (value function yet not available for Mark-III); the European EuroQol-5D instrument; and the Quality of Well-Being scale developed in the United States.

Based on the results obtained with its predecessor the HSCS, the Health Utility Index Mark-II (HUI Mark-II) is currently probably the most comprehensive and widely documented classification system (Torrance *et al.*, 1996). Furthermore, it is the only system hitherto that is intentionally based on multiattribute utility theory (MAUT). Its classification is focused on a functional concept of health status, though focused on disabilities (dysfunction) and resulting dependencies measured by seven attributes. The KR-multiplicative HUI Mark-II function is:

$$U(Q) = c\,(w_1q_1 \times w_2q_2 \times w_3q_3 \times w_4q_4 \times w_5q_5 \times w_6q_6 \times w_7q_7) + (1-c). \qquad (7)$$

The next two classification systems were both not explicitly developed from MAUT, but can appropriately be regarded as additive value functions. The EQ-5D measure is the instrument which, compared to the other two classification systems, defines health status most in terms of health-related quality-of-life. The initial (unofficial) function presented by the EuroQol Group for their EuroQol multiattribute instrument was based on the additive function (Hout van & McDonnell, 1993):

(8)

$$U(Q) = c + w_1q_1 + w_2q_2 + w_3q_3 + w_4q_4 + w_5q_5.$$

In 1997 other researchers of the EuroQol Group published another "tariff" for their EuroQol-5D instrument based on an additive model extended with one "global" interaction term (Dolan, 1997) and additional parameters to adjust for the non-equidistance between level 1 to level 2 and level 2 to level 3. The interaction term, called the N3 term, is based on a more extended modeling model (random effects model). Some questions about the

assumptions that have been made (redundancy of parameters) and data transformations (individual rescaling of "death" values) that were performed have yet to be answered.

$$U(Q) = c^{*} - (w_1 q_1^{\,i} + w_2 q_2^{\,i} + w_3 q_3^{\,i} + w_4 q_4^{\,i} + w_5 q_5^{\,i} + w_6 q_1^{\,ii} + w_7 q_2^{\,ii} + w_8 q_3^{\,ii} + w_9 q_4^{\,ii} + w_{10} q_5^{\,ii} + w_{11} q^{***}) \quad (9)$$

Parameter c^{*} in equation (9) represents a constant term that is included for each dysfunctional state (e.g., worse than health state "11111"). Parameter w_{11} represents a parameter that is included if there occurs a level 3 within at least one dimension (q^{**}). So far, the advantage of model (9) above model (8) is unclear.

A third health-status classification system, the American Quality of Well-Being scale is based on a straightforward additive function without terms to adjust for possible interactions between the four attributes and their levels. The health-status concept of this instrument is more properly indicated as a health symptom-problem concept, a mix of functional health status like HUI and one symptom-specific attribute.

$$U(Q) = c + (w_1 q_1 + w_2 q_2 + w_3 q_3 + w_4 q_4) \quad (10)$$

Discussion

Three important aspects can be noticed in regard to the construction of multiattribute classification systems. First, the choice of a health concept determines the validity of the classification system. For the use of the QALY concept, it is becoming a convention to use values that reflect the health-related quality-of-life, for which we have used the term health status. A classification system is therefore more adequate if it contains several domains (attributes) related to health. The tradeoff between the number of relevant domains and the number of attributes of a classification system is a difficult one. Inevitably, the number of attributes has to be limited, not only because of modeling consequences, but also because the number of attributes human beings can cognitively process is restricted. The cognitive psychologist Miller (1956) concluded after several experiments: "There seems to be some limitation built into us either by learning or by the design of our nervous systems, a limit that keeps our channel capacities (information process, author) in this general range (5 – 9, author)".

Second, the number of levels determines the precision of a multiattribute classification system. In combination with the modeling strategy conducted, the number of levels merely affects the reliability of the values elicited by classification systems.

Third, the choice for a specific model (value function) is another important aspect that greatly determines the reliability of the classification system. However, if modeling is performed based on inadequate theoretical assumptions or if the modeling is conducted in a straightforward manner, but not in a way that fully explains the underlying relationships

between the different attributes and their combinations of levels, this poses a further threat to the validity of the values of such classification systems.

Clearly, this area needs further development, starting with a multi-disciplinary language to describe and relate the different approaches. We hope that Chapter 2 and the Appendices A–C may serve that goal.

References

Abdalla M, Russell I (1995) *Tariffs for the EuroQol health states based on modelling the individual VAS and TTO data of the York survey.* In: O'Hanlo M, Buxton M, eds. Conference proceedings of the EuroQol Plenary Meeting 1994. London, Discussion Papers, HERG Research Report No. 20: 75-91.

Agt HME van, Essink-Bot ML, Krabbe PFM, Bonsel GJ (1994) Test-retest reliability of health state valuations collected with the EuroQol questionnaire. *Social Science & Medicine* 11: 1537-1544.

Aitken RCB (1969) Measurement of feelings using visual analogue scales. *Proceedings of the Royal Society of Medicine* 62: 989-993.

Anderson NH (1970) Functional measurement and psycho-physical judgment. *Psychological Review* 77: 153-169.

Anderson NH (1971) Integration theory and attitude change. *Psychological Review* 78: 171-206.

Anderson NH (1976) How functional measurement can yield validated interval scales of mental quantities. *Journal of Applied Psychology* 61: 677-692.

Aristoteles (1954) *Ethica Nicomachea,* Antwerpen: De Nederlandse boekhandel. [originally published around 345 B.C.]

Barendregt JJ, Bonneux L, Maas PJ van der (1996) DALYs: the age-weights on balance. *Bulletin of the World Health Organization* 74: 439-443.

Bass EB, Steinberg EP, Pitt HA, Griffiths RI, Lillemoe KD, Saba GP, Johns C (1994) Comparison of the rating scale and the standard gamble in measuring patient preferences for outcomes of gallstone disease. *Medical Decision Making* 14: 307-314.

Bleichrodt H (1996) *Applications of utility theory in the economic evaluation of health care.* Thesis Erasmus University Rotterdam, Rotterdam.

Bleichrodt H, Johannesson M (1997) The validity of QALYs: an experimental test of constant proportional trade-off and utility independence. *Medical Decision Making* 17: 21-32.

BMDP Software (1992) *BMDP Statistical Software Manual, Vol. 2, release 7.0,* Berkeley: University of California Press.

Bogaert W van den, Ostyn F, Schueren E van der (1983) The primary treatment of advanced vocal cord cancer: laryngectomy or radiotherapy? *International Journal of Radiation Oncology Biology* 9: 329-334.

Bombardier C, Wolfson AD, Sinclair AJ, McGreer A (1982) Comparison of three preference measurement methodologies in the evaluation of a functional index. In: Deber R, Thompson G, eds. *Choices in Health Care: Decision Making and Evaluation of Effectiveness.* Toronto: University of Toronto.

Bonsel GJ (1991) *Methods of medical technology assessment with an application of liver transplantation.* Thesis Erasmus University Rotterdam, Rotterdam.

Bonsel GJ, Agt HME van (1994) *The number of levels in the descriptive system.* In: Busschbach JJ van, Bonsel GJ, Charro FTh de, eds. Conference proceedings of the EuroQol Plenary Meeting 1993. Rotterdam, Discussion Papers, Department of Public Health: 115-121.

Bonsel GJ, Essink-Bot ML, Klompmaker IJ, Slooff MJH (1992) Assessment of the quality of life before and following liver transplantation. *Transplantation* 53: 796-800.

Bonsel GJ, Klompmaker IJ, Essink-Bot ML, Habbema JDF, Slooff MJH (1990) Cost-effectiveness analysis of the Dutch liver transplantation programme. *Transplantation Proceedings* 22: 1481-1484.

Bosch JL, Hunink MGM (1996) The relationship between descriptive and valuational quality-of-life measures in patients with intermittent claudication. *Medical Decision Making* 16: 217-225.

Bossuyt P (1990) *A comparison of probabilistic unfolding theories for paired comparisons.* Berlin: Springer-Verlag.

Boyd NF, Sutherland HJ, Heasman KZ, Tritchler DL (1990) Whose utilities for decision analysis? *Medical Decision Making* 10: 58-67.

Brooks RG (1995) *Health status measurement: a perspective on change*, London: Macmillan Press.

Brooks RG (1996) EuroQol: the current state of play. *Health Policy* 37: 53-72.

Brooks RG, Jendteg S, Lindgren B, Persson U, Björk S (1991) EuroQol: health-related quality of life measurement. Results of the Swedish questionnaire exercise. *Health Policy* 18: 37-48.

Broome J (1993) QALYs. *Journal of Public Economics* 50: 149-167.

Bühler K (1929) *Die Krise der Psychologie.* Jena: Verlag von Gustav Fischer.

Busschbach JJ van, Hessing DJ, Charro FTh de (1994) *An empirical comparison of four measurements of quality of life: standard gamble, time trade-off, the EuroQol-visual analogue scale and the Rosser & Kind matrix.* In: Sintonen H, ed. Conference proceedings of the EuroQol Plenary Meeting 1993. Discussion Papers, Helsinki, Kuopio University Publications: 41-53.

Busschbach JJ van, Hessing DJ, Charro FTh de (1993) The utility of health at different stages in life: a quantitative approach. *Social Science & Medicine* 37: 153-158.

Busschbach JJ van, McDonnell J, Hout BA van (1997) *Testing different parametric relation between the EuroQol health description and health valuation in students.* In: Nord E, ed. Conference proceedings of the EuroQol Plenary Meeting 1996. Oslo, National Institute of Public Health.

Cadman D, Goldsmith C (1986) Construction of social value or utility-based health indices: the usefulness of factorial experimental designs. *Journal of Chronic Diseases* 39: 643-651.

Campbell DT, Fiske DW (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56: 81-105.

Carr-Hill R (1989) Assumptions of the QALY procedure. *Social Science & Medicine* 29: 469-477.

Chapman GB, Elstein AS (1995) Valuing the future: temporal discounting of health and money. *Medical Decision Making* 15: 373-386.

Chapman GB, Elstein AS, Hughes KK (1995) Effects of patient education on decisions about breast cancer treatments: a preliminary report. *Medical Decision Making* 15: 231-239.

Chestnut LG, Keller LR, Lambert WE, Rowe RD (1996) Measuring heart patient' willingness to pay for changes in angina symptoms. *Medical Decision making* 16: 65-77.

Christensen-Szalanski JJJ (1984) Discount functions and the measurement of patients' values: women's decisions during childbirth. *Medical Decision Making* 4: 47-58.
Cohen BJ (1996a) Assigning values to intermediate health states for cost-utility analysis. *Medical Decision Making* 16: 376-385.
Cohen BJ (1996b) Is expected utility theory normative for medical decision making? *Medical Decision making* 16: 1-6.
Cohen J (1977) *Statistical power analyisis for the behavioral sciences*, New York: Academic Press.
Coombs CH (1950) Psychological scaling without a unit of measurement. *Psychological Review* 57: 145-158.
Coombs CH (1964) *A theory of data*, New York: John Wiley.
Crocker L, Algina J (1986) *Introduction to classical and modern test theory*, New York: Holt, Rinehart and Winston.
Cronbach LJ, Rajaratnam N, Gleser GC (1963) Theory of generalizability: a liberation of reliability theory. *The Britsh Journal of Statistical Psychology* 16: 132-163.
David HA (1988) *The method of paired comparisons*, Charles Griffin & Company Limited, London.
Descartes R (1969) *Discours de la Méthode*, Hamburg: Felix Meiner Verlag. [originally published in 1673]
Deyo RA, Diehr P, Patrick DL (1991) Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. *Controlled Clininal Trials* 12: 142S-158S.
Dolan P (1997) Modeling valuations for EuroQol health states. *Medical Care* 35: 1095-1108.
Dolan P, Gudex C (1995) Time preference, duration and health state valuations. *Health Economics* 4: 289-299.
Dolan P, Gudex C, Kind P, Wiliams A (1996) The time trade-off method: results from a general population study. *Health Economics* 5: 141-154.
Donaldson C, Shackley P (1997) Does "process utility" exist? A case study of willingness to pay for laparoscopic cholecystectomy. *Social Science & Medicine* 44: 699-707.
Donk J van der, Levendag PC, Kuijpers AJ, Roest FHJ, Habbema JDF, Meeuwis CA, Schmitz PIM (1995) Patients' participation in clinical decision-making for treatment of T3 laryngeal cancer: a comparison of state and process utilities. *Journal of Clinical Oncology* 9: 2369-2378.
Drummond MF, O'Brien B, Stoddart GL, Torrance GW (1997) *Methods for the economic evaluation of health care programmes*, Oxford: Oxford University Press.
Duncan C, Jones K, Moon G (1998) Context, composition and heterogeneity: using multilevel models in health research. *Social Science & Medicine* 46: 97-117.
Elsinga E, Rutten FFH (1997) Economic evalution in support of national health policy: the case of the Netherlands. *Social Science & Medicine* 45: 605-620.
Elstein AS (1989) Decision analysis in surgical education. *World Journal of Surgery* 13: 287-291.
Elstein AS, Shulman LS, Sprafka SA (1978) *Medical problem solving: an analysis of clinical reasoning*, Cambridge: Harvard University Press.
Eraker SA, Polister P (1982) How decisions are reached: physician and patient. *Annals of Internal Medicine* 97: 262-268.

Essink-Bot ML (1995) *Health status as a measure of outcome of disease and treatment.* Thesis Erasmus University Rotterdam, Rotterdam.

Essink-Bot ML, Bonsel GJ, Maas PJ van der (1990) Valuation of health states by the general public: feasibility of a standardized measurement procedure. *Social Science & Medicine* 31: 1201-1206.

Essink-Bot ML, Krabbe PFM, Bonsel GJ, Aaronson NK (1997) An empirical comparison of four generic health status measures: the Nottingham Health Profile, the MOS 36-Item Short-Form Health Survey, the COOP/WONCA charts, and the EuroQol instrument. *Medical Care* 35: 522-537.

Essink-Bot ML, Stouthard MEA, Bonsel GJ (1993) Generalizability of valuations on health states collected with the EuroQol-questionnaire. *Health Economics* 2: 237-246.

EuroQol Group (1990) EuroQol – a new facility for the measurement of health-related quality of life. *Health Policy* 16: 199-208.

Fanshel S, Bush JW (1970) A health-status index and its application to health-services outcomes. *Operations Research* 18: 1021-1066.

Feeny D, Furlong W, Barr RD, Torrance GW, Rosenbaum P, Weitzman S. (1992) A comprehensive multiattribute system for classifying the health status of survivors of childhood cancer. *Journal of Clinical Oncology* 10: 923-928.

Feeny D, Torrance G, Goldsmith C, Furlong W, Boyle M (1994) *A multi-attribute approach to population health status.* working paper 94-5, McMaster University Centre for Health Economics and Policy Analysis Hamilton, Ontario.

Fischer GW (1979) Utility models for multiple objective decisions: do they accurately represent human preferences? *Decision Sciences* 10: 451-479.

Froberg DG, Kane RL (1989a) Methodology for measuring health-state preferences - I: measurement strategies. *Journal of Clinical Epidemiology* 42: 345-354.

Froberg DG, Kane RL (1989b) Methodology for measuring health-state preferences - II: scaling methods. *Journal of Clinical Epidemiology* 42: 459-471.

Froberg DG, Kane RL (1989c) Methodology for measuring health-state preferences - III: population and context effects. *Journal of Clinical Epidemiology* 42: 585-592.

Froberg DG, Kane RL (1989d) methodology for measuring health-state preferences - IV: progress and a research agenda. *Journal of Clinical Epidemiology* 42: 675-685.

Fryback DG, Dasbach EJ, Klein R, Klein BEK, Dorn N (1993) The Beaver Dam health outcomes study: initial catalog of health-state quality factors. *Medical Decision Making* 13: 89-102.

Fryback DG, Lawrence WF (1997) Dollars may not buy as many QALYs as we think: a problem with defining quality-of-life adjustments. *Medical Decision Making* 17: 276-284.

Gafni A (1991) Willingness-to-pay as a measure of benefits: relevant questions in the context of public decisionmaking about health care programs. *Medical Care* 29: 1246-1252.

Gafni A (1994) The standard gamble method: what is being measured and how it is interpreted. *Health Services Research* 29: 207-224.

Gafni A (1995) Time in health: can we measure individuals' "pure time preferences"? *Medical Decision Making* 15: 31-37.

Gafni A (1996) HYEs: do we need them and can they fulfil the promise? *Medical Decision Making* 16: 215-216.

Gafni A, Torrance GW (1984) Risk attitude and time preference in health. *Management Science* 30: 440-451.

Goethe JW von (1981) *Werke in zwei Bänden* [I] / Maximen und Reflexionen, München: Carl Hanser Verlag. [originally published in the journal "Kunst und Altertum" between 1818-1827]

Gold MR, Siegel JE, Russel LB, Weinstein MC (1996) *Cost-effectiveness in health and medicine*, New York: Oxford University Press.

Goldstein H (1995) *Multilevel statistical models*, New York: Edward Arnold.

Green PE, Rao VR (1971) Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research* 8: 355-363.

Grogono AW, Woodgate DJ (1971) Index for measuring health. *The Lancet* 297: 1024-1026.

Hadorn DC, Hays RD (1991) Multitrait-multimethod analysis of health-related quality-of-life measures. *Medical Care* 29: 829-840.

Hadorn DC, Hays RD, Uebersax J, Hauber T (1992) Improving task comprehension in the measurement of health state preferences: a trial of informational cartoon figures and a paired-comparison task. *Journal of Clinical Epidemiology* 45: 233-243.

Hakim Z, Pathak DS, Kucukarslan S, Schommer JC (1995) Comparing alternative methodologies for measuring health state preferences. *MAPI Quality of Life News Letter* 12: 2/6.

Harris J (1987) QALYfying the value of life. *Journal of Medical Ethics* 13: 117-123.

Harris J (1988) Life: quality, value and justice. *Health Policy* 10: 259-266.

Hayes MHS, Patterson DG (1921) Experimental development of the graphic rating method. *American Psychological Association* 18: 98-99.

Hershey JC, Kunreuther HC, Schoemaker PJH (1982) Sources of bias in assessment procedures for utility functions. *Management Science* 28: 936-954.

Hilden J (1985) The nonexistence of interpersonal utility scales: a missing link in medical decision theory? *Medical Decision Making* 5: 215-228.

Hilden J, Glasziou PP, Habbema JDF (1992) A pitfall in utility assessment - patients' undisclosed investment decisions. *Medical Decision Making* 12: 39-43.

Holmes AM (1995) A QALY-based societal health statistic for Canada, 1985. *Social Science & Medicine* 41: 1417-1427.

Hornberger JC, Redelmeier DA, Petersen J (1992) Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis. *Journal of Clinical Epidemiology* 45: 505-512.

Hout BA van, Bonsel GJ, Habbema JDF, Maas PJ van der, Charro FTh de (1993a) Heart transplantation in the Netherlands; costs, effects and scenarios. *Journal of Health Economics* 12: 73-93.

Hout BA van, McDonnell J (1993) *Estimating a parametric relation between health description and health valuation using the EuroQol instrument.* In: Björk S, ed. Conference proceedings of the EuroQol Plenary Meeting 1992. IHE Working Paper, Lund, Sweden, The Swedish Institute for Health Economics: 45-59.

Hout BA van, Wielink G, Bonsel GJ, Rutten FFH (1993b) Effects of ACE inhibitors on heart failure in the Netherlands. A pharmacoeconomic model. *Pharmacoeconomics* 3: 387-397.

Johannesson M (1995) The ranking properties of healthy-years equivalents and quality-adjusted life-years under certainty and uncertainty. *International Journal of Technology Assessment in Health Care* 11: 40-48.

Johannesson M, Jönsson B, Borgquist L (1991) Willingness to pay for antihypertensive therapy - results of a Swedish pilot study. *Journal of Health Economics* 10: 461-474.

Johannesson M, Jönsson B, Karlsson G (1996) Outcome measurement in economic evaluation. *Health Economics* 5: 279-296.

Johannesson M, Pliskin JS, Weinstein MC (1994) A note on QALYs, time tradeoff, and discounting. *Medical Decision Making* 14: 188-193.

Jöreskog KG, Sörbom D (1979) *Advanges in factor analysis and structural equation models*, Cambridge: Abt Books.

Jöreskog KG, Sörbom D (1989) *LISREL 7: a guide to the program and applications*, Chicago: SPSS.

Kahneman D, Tversky A (1982) The psychology of preferences. *Scientific America* 246: 160-173.

Kaplan RM, Anderson JP (1988) A general health policy model: update and application. *Health Services Research* 23: 199-235.

Kaplan RM, Bush JW, Berry CC (1976) Health status: types of validity and the index of well-being. *Health Services Research* 11: 478-507.

Kaplan RM, Feeny D, Revicki DA (1993) Methods for assessing relative importance in preference based outcome measures. *Quality of Life Research* 2: 467-475.

Kartman B, Andersson F, Johannesson M (1996) Willingness to pay for reductions in angina pectoris attacks. *Medical Decision Making* 16: 248-253.

Kassirer JP (1983) Adding insult to injury: usurping patients' prerogatives. *The New England Journal of Medicine* 308: 898-901.

Keeney RL, Raiffa H (1976) *Decisions with multiple objectives: preferences and value tradeoffs*, New York: John Wiley & Sons.

Kerlinger FN (1986) *Foundations of behavioral research*, New York: CBS Publishing.

Kind P, Leese B, Hardman G (1998) *Demonstrating the value of EQ-5D as a tool for monitoring health status in primary care.* In: Busschbach van JJ, ed. Conference proceedings of the EuroQol Plenary Meeting 1997. Rotterdam. [in press]

Krabbe PFM, Bonsel GJ (1998) Sequence effects, health profiles and the QALY model: in search of realistic modeling. *Medical Decision Making* 18: 178-186.

Krabbe PFM, Bonsel GJ, Essink-Bot ML (1995) *Modelling: multiple classification analyse and logit transformation.* In: O'Hanlon M, Buxton M, eds. Conference proceedings of the EuroQol Plenary Meeting 1994. London, Discussion Papers: Brunel University: 93-107.

Krabbe PFM, Essink-Bot ML, Bonsel GJ (1996) On the equivalence of collectively and individually collected responses: standard-gamble and time-tradeoff judgments of health states. *Medical Decision Making* 16: 120-132.

Krabbe PFM, Essink-Bot ML, Bonsel GJ (1997) The comparability and reliability of five health-state valuation methods. *Social Science & Medicine* 45: 1641-1652.

Kruskal JB, Wish M (1978) *Multidimensional scaling*, Beverly Hills and London: Sage Publications.

Kupperman M, Shiboski S, Feeny D, Elkin EP, Washington AE (1997) Can preference scores for discrete states be used to derive preference scores for an entire path of events? An application of prenatal diagnosis. *Medical Decision Making* 17: 42-55.

La Puma J, Lawlor EF (1990) Quality-adjusted life-years: ethical implications for physicians and policymakers. *Journal of the American Medical Association* 263: 2917-2921.

Lee PR (1967) Health and well-being. *The Annals of the American Academy*: 193-207.

Levendag PC, Ravasz LA, Terhaard CHJ, Hordijk GJ (1993) T3 squamous cell carcinoma of the larynx treated by a split-course radiation protocol. *American Journal of Clinical Oncology* 16: 509-518.

Lewis-Beck MS (1995) *Basic measurement*, London: Sage Publications.

Lipscomb J (1989) Time preference for health in cost-effectiveness analysis. *Medical Care* 27: S233-S253.

Llewellyn-Thomas H, Sutherland HJ, Tibshirani R (1982) The measurement of patients' values in medicine. *Medical Decision Making* 2: 449-462.

Llewellyn-Thomas H, Sutherland HJ, Tibshirani R (1984a) Describing health states: methodological issues in obtaining values for health states. *Medical Care* 22: 543-552.

Llewellyn-Thomas HA, Sutherland HJ, Ciampi A (1984b) The assessment of values in laryngeal cancer: reliability of measurement methods. *Journal of Chronic Diseases* 37: 283-291.

Loewenstein GF, Prelec D (1993) Preferences for sequences of outcomes. *Psychological Review* 100: 91-108.

Loomes G (1993) Disparities between health state measures: is there a rational explanation? In: Gerrard B ed. *The economics of rationality*. London, Routledge: 149-178.

Loomes G, McKenzie L (1989) The use of QALYs in health care decision making. *Social Science & Medicine* 28: 299-305.

Louviere JJ (1988) Conjoint analysis modelling of stated preferences: a review of theory, methods, recent developments and external validity. *Journal of Transport Economics and Policy* 22: 93-119.

Luce RD, Tukey JW (1964) Simultaneous conjoint measurement: a new type of fundamental measurement. *Journal of Mathematical Psychology* 1: 1-27.

Maas A, Stalpers L (1992) Assessing utilities by means of conjoint measurement: an application in medical decision analysis. *Medical Decision Making* 12: 288-297.

Maas A, Wakker P (1994) Additive conjoint measurement for multiattribute utility. *Journal of Mathematical Psychology* 38: 86-101.

MacCallum RC (1978) Recovery of structure in incomplete data by ALSCAL. *Psychometrika* 44: 69-74.

Mayo NE, Goldberg M, Kind P (1998) *Calibration of the EuroQol EQ-5D in a Canadian population survey: initial results from the CHIC study*. In: Busschbach van JJ, ed. Conference proceedings of the EuroQol Plenary Meeting 1997. Rotterdam. [in press]

McIver JP, Carmines EG (1981) *Unidimensional scaling*, London: Sage Publications.

McNeil BJ, Weichselbaum R, Pauker SG (1981) Speech and survival: tradeoffs between quality and quantity of life in laryngeal cancer. *The New England Journal of Medicine* 305: 982-987.

Mehrez A, Gafni A (1989) Quality-adjusted life years, utility theory, and healthy-years equivalents. *Medical Decision Making* 9: 142-149.

Mehrez A, Gafni A (1991) The healthy-years equivalents: how to measure them using the standard gamble approach. *Medical Decision Making* 11: 140-146.

Mehrez A, Gafni A (1993) Healthy-years equivalents versus quality-adjusted life years: in pursuit of progress. *Medical Decision Making* 13: 287-292.

Miller GA (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* 63: 81-97.

Miyamoto JM, Eraker SA (1985) Parameter estimates for a QALY utility model. *Medical Decision Making* 5: 191-213.

Murray CJL, Lopez AD (1994) Quantifying the burden of disease: the technical basis for disability-adjusted life years. In: Murray CJL, Lopez AD eds. *Global comparative assessments in the health sector: disease burden, expenditures and intervention packages.* Geneva, World Health Organization: 3-19.

Murray CJL, Lopez AD (1996) *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020.* Global burden of disease and injury series. Geneva: Harvard School of Public Health/World Health Organization.

Murray CJL, Lopez AD (1997a) Alternative projections of mortality and disability by cause 1990-2020 Global Burden of Disease Study. *The Lancet* 349: 1498-1504.

Murray CJL, Lopez AD (1997b) Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study. *The Lancet* 349: 1436-1442.

Murray CJL, Lopez AD (1997c) Mortality by cause for eight regions of the world: Global Burden of Disease Study. *The Lancet* 349: 1269-1276.

Murray CJL, Lopez AD (1997d) Regional patterns of disability-free life expectancy and disability-adjusted life expectancy: Global Burden of Disease Study. *The Lancet* 349: 1347-1352.

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

Nease RF (1997) Do violations of the axioms of expected utility theory threaten decision analysis? *Medical Decision Making* 16: 399-403.

Neumann J von, Morgenstern O (1953) *Theory of games and economic behavior*, Princeton: Princeton University Press.

Nord E (1991) EuroQol: health-related quality of life measurement. Valuations of health states by the general public in Norway. *Health Policy* 18: 25-36.

Nord E (1992) Methods for quality adjustment of life years. *Social Science & Medicine* 34: 559-569.

Nord E (1995) The person-trade-off approach to valuing health care programs. *Medical Decision Making* 15: 201-208.

Nord E, Richardson J, Street A, Kuhse H, Singer P (1995) Maximizing health benefits vs egalitarianism: an Australian survey of health issues. *Social Science & Medicine* 41: 1429-1437.

O'Brien B, Gafni A (1996) When do the "dollars" make sense? Toward a conceptual framework for contingent valuation studies in health care. *Medical Decision Making* 16: 288-299.

O'Brien B, Viramontes JL (1994) Willingness to pay: a valid and reliable measure of health state preference? *Medical Decision Making* 14: 289-297.

O'Connor AM (1989) Effects of framing and level of probability on patients' preferences for cancer chemotherapy. *Journal of Clinical Epidemiology* 42: 119-126.

O'Connor AMC, Boyd NF, Warde P, Stolbach L, Till JE (1987) Eliciting preferences for alternative drug therapies in oncology: influence of treatment outcome description, elicitation technique and treatment experience on preferences. *Journal of Chronic Diseases* 40: 811-818.

O'Hanlon M, Fox-Rushby J, Buxton MJ (1994) A qualitative and quantitative comparison of the EuroQol and time trade-off techniques. *International Journal of Health Sciences* 5: 85-97.

Ohinma A, Helala E, Sintonen H (1996) *Modelling EuroQol values of Finnish adult population.* In: Badia X, Herdman M, Segura A, eds. Conference proceedings of the EuroQol Plenary Meeting 1995. Barcelona, Discussion Papers: Institut Universitari de Salut Pública de Catalunya: 67-76.

Patrick DL, Bush JW, Chen MM (1973a) Methods for measuring levels of well-being for a health status index. *Health Services Research* 8: 228-245.

Patrick DL, Bush JW, Chen MM (1973b) Toward an operational definition of health. *Journal of Health & Social Behavior* 14: 6-23.

Patrick DL, Starks HE, Cain KC, Uhlmann RF, Pearlman R (1994) Measuring preferences for health states worse than death. *Medical Decision Making* 14: 9-18.

Pauker SG, Kassirer JP (1987) Decision analysis. *The New England Journal of Medicine* 316: 250-258.

Pauker SG, McNeil BJ (1981) Impact of patient preferences on the selection of therapy. *Journal of Chronic Diseases* 34: 77-86.

Payne JW (1982) Contingent decision behavior. *Psychological Bulletin* 92: 382-402.

Pliskin JS, Beck CH (1976) A health index for patient selection: a value function approach with application to chronic renal failure patients. *Management Science* 22: 1009-1021.

Pliskin JS, Shepard DS, Weinstein MC (1980) Utility functions for life years and health status. *Operations Research* 28: 206-224.

Rawls J (1997) *A theory of justice*, Oxford: Oxford University Press. [originally published in 1971]

Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC (1984) Preferences for health outcomes: comparison of assessment methods. *Medical Decision Making* 4: 315-329.

Redelmeier DA, Heller DN (1993) Time preference in medical decision making and cost-effectiveness analysis. *Medical Decision Making* 13: 212-217.

Richards MA, Ramirez AJ, Degner LF, Fallowfield LJ, Mahler EJ, Neuberger J (1995) Offering choise of treatment to patients with cancers: a review based on a symposium held at the 10th annual conference of the British Psychological Oncology Group, december 1993. *European Journal of Cancer* 31A: 112-116.

Richardson J (1994) Cost utility analysis: what should be measured? *Social Science & Medicine* 39: 7-21.

Richardson J, Hall J, Salkfeld G (1996) Measurement of utility in multiphase health states. *International Journal of Technology Assessment in Health Care* 12: 151-162.

Roest FHJ, Eijkemans MJC, Donk J van der, Levendag PC, Meeuwis CA, Schmitz PIM, Habbema JDF (1997) The use of confidence intervals for individual utilities: limits to formal decision analysis for treatment choice. *Medical Decision Making* 17: 285-291.

Rosser R, Kind P (1978) A scale of valuations of states of illness: is there a social consensus? *International Journal of Epidemiology* 7: 347-358.

Ruwaard D, Kramers PGN (1997) *Volkgezondheid Toekomst Verkenningen: de som der delen.* Bilthoven: Rijksinstituut voor Volkgezondheid en Milieu. [in Dutch]

Schmitt N, Stults DM (1986) Methodology review: analysis of multitrait-multimethod matrices. *Applied Psychological Measurement* 10: 1-22.

Schoemaker PJH (1982) The expected utility model: its variants, purposes, evidence and limitations. *Journal of Economic Literature* 20: 529-563.

Selai CE (1996) *Four levels and a funeral: prelimanary testing of a four-level EuroQol descriptive classification system.* In: Badia X, Herdman M, Segura A, eds. Conference proceedings of the EuroQol Plenary Meeting 1995. Barcelona, Discussion Papers: Institut Universitari de Salut Pública de Catalunya: 33-43.

Selai CE, Rosser RM (1993) Good quality quality? Some methodological issues. *Journal of the Royal Society of Medicine* 86: 440-443.

Selai CE, Rosser RM (1995) Eliciting EuroQol descriptive data and utility scale values from inpatients. *PharmacoEconomics* 8: 147-158.

Shavelson RJ, Webb NM, Rowley GL (1989) Generalizability theory. *American Psychologist* 44: 922-932.

Shepard RN (1981) Psychological relations and psychophysical scales: on the status of "direct" psychophysical measurement. *Journal of Mathematical Psychology* 24: 21-57.

Siegel S, Castellan NJ (1988) *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill Book Company.

Sintonen H (1981) An approach to measuring and valuing health status. *Social Science & Medicine* 15C: 55-65.

Smith A (1987) Qualms about QALYs. *The Lancet* 329: 1134-1136.

Stålhammar NO (1996) An empirical note on willingness to pay and starting-point bias. *Medical Decision Making* 16: 242-247.

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

Staquet MJ, Hays RD, Fayers PM (1998) *Quality of life assessment in clinical trials*, Oxford: Oxford University Press.

Stevens SS (1946) On the theory of scales of measurement. *Science* 103: 677-680.

Stevens SS (1957) On the psychophysical law. *The Psychological Review* 64: 153-181.

Stiggelbout AM, Eijkemans MJC, Kiebert GM, Kievit J, Leer JWH, Haes HJCJM de (1996) The "utility" of the visual anolog scale in medical decision making and technology assessment: is it an alternative to the time trade-off? *International Journal of Technology Assessment in Health Care* 12: 291-298.

Stiggelbout AM, Kiebert GM, Kievit J, Leer JWH, Stoter G (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.

Stouthard MEA, Essink-Bot ML, Bonsel GJ, Barendregt JJ, Kramer PGN, Water HPA van de, Gunning-Schepers LJ, Maas PJ van der (1997) *Disability weights for diseases in the Netherlands*. Rotterdam, Erasmus University Rotterdam: Department of Public Health.

Streiner DL, Norman GR (1995) *Health measurement scales: a practical guide to their development and use*, Oxford: Oxford University Press.

Stronks K, Strijbis AM, Wendte JF, Gunning-Schepers LG (1997) Who should decide? Qualitative analysis of panel data from public, patients, healthcare professionals, and insurers on priorities in health care. *British Medical Journal* 315: 92-96.

Sutherland HJ, Dunn V, Boyd NF (1983) Measurement of values for states of health with linear analog scales. *Medical Decision Making* 3: 477-487.

Sutherland HJ, Llewellyn-Thomas H, Boyd NF, Till JE (1982) Attitudes toward quality of survival: the concept of "maximal endurable time". *Medical Decision Making* 2: 299-309.

Thompson MS (1986) Willingness to pay and accept risks to cure chronic disease. *American Journal of Public Health* 76: 392-396.

Thompson MS, Read JL, Liang M (1982) Willingness-to-pay concepts for societal decisions in health. In: Kane RL, Kane RA eds. *Values and long-term care*. Lexington, University of California: 103-125.

Thompson MS, Read JL, Liang M (1984) Feasibility of willingness-to-pay measurement in chronic arthritis. *Medical Decision Making* 4: 195-215.

Torgerson WS (1958) *Theory and methods of scaling*, New York: Wiley.

Torrance GW (1976) Social preferences for health states: an empirical evaluation of three measurement techniques. *Socio-Economic Planning Sciences* 10: 129-136.

Torrance GW (1982a) Multiattribute utility theory as a method of measuring social preferences for health states in long-term care. In: Kane RL, Kane RA eds. *Values and long-term care*. Lexington, University of California: 127-155.

Torrrance GW (1982b) Preferences for health states: a review of measurement methods. *Mead Johson Symposium on Perinatal & Development Medicine*, 20: 37-45.

Torrance GW (1986) Measurement of health state utilities for economic appraisal: a review. *Journal of Health Economics* 5: 1-30.

Torrance GW (1987) Utility approach to measuring health-related quality of life. *Journal of Chronical Diseases* 40: 593-600.

Torrance GW, Boyle MH, Horwood SP (1982) Application of multi-attribute utility theory to measure social preferences for health states. *Operations Research* 30: 1043-1069.

Torrance GW, Feeny D (1989) Utilities and quality-adjusted life years. *International Journal of Technology Assessment in Health Care* 5: 559-575.

Torrance GW, Feeny D, Furlong WJ, Barr RD, Zhang YZ, Wang Q (1996) Multiattribute utility function for a comphrensive health status classsification system: Health Utility Index Mark 2. *Medical Care* 34: 702-722.

Torrance GW, Furlong W, Feeny D, Boyle M (1995) Multi-attribute preference functions: health utility index. *PharmacoEconomics* 7: 503-520.

Torrance GW, Thomas WH, Sackett DL (1972) A utility maximization model for evaluation of health care programs. *Health Services Research* 3: 118-133.

Torrance GW, Zhang Y, Feeny D, Furlong W, Barr R (1992) *Multi-attribute preference functions for a comprehensive health status classification system.* Working Paper 92-18, Hamilton, Ontario, McMaster University, Health Sciences Centre Hamilton.

Tversky A, Kahneman D (1974) Judgement under uncertainty: heuristics and biases. *Science* 185: 1124-1131.

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

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

Veit CT, Rose BJ, Ware JE (1982) Effects of physical and mental health on health-state preferences. *Medical Care* 20: 386-401.

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

Vroon PA (1972) *Enkele psychofysische en cognitieve aspecten van de tijdzin* [some psychophysical and cognitive aspects of the time sense]. Thesis University of Utrecht, Utrecht. [with a summary in English]

Wakker P (1996) A criticism of healthy-years equivalents. *Medical Decision Making* 16: 207-214.

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

Weinstein MC (1993) Time preference studies in the health care context. *Medical Decision Making* 13: 218-219.

Weinstein MC, Fineberg HV, Elstein AS, Frazier HS, Neuhauser D, Neutra RR, McNeil BJ (1980) *Clinical decision analysis.* Philadelphia: Saunders.

Weinstein MC, Pliskin JS (1996) Perspectives on healthy-years equivalents HYEs. What are the issues. *Medical Decision Making* 16: 205-206.

Weinstein MC, Siegel JE, Gold MR, Kamlet MS, Russell LB (1996) Recommendations of the panel on cost-effectiveness in health and medicine. *Journal of the American Medical Association* 276: 1253-1258.

Weinstein MC, Stason WB (1977) Foundations of cost-effectiveness analysis for health and medical practices. *The New England Journal of Medicine* 296: 716-721.

Wewers ME, Lowe NK (1990) A critical review of visual analogue scales in the measurement of clinical phenomena. *Research in Nursing & Health* 13: 227-236.

Williams A (1985) Economics of coronary artery bypass grafting. *British Medical Journal* 291: 326-329.

Williams A (1995) *The role of the EuroQol instrument in QALY calculations.* Discussion Paper 130, University of York: Centre for Health Economics.

Williams A (1996) QALYs and ethics: a health economist's perspective. *Social Science & Medicine* 43: 1705-1804.

Williams A (1997) Intergenerational equity: an exploration of the "fair innings" argument. *Health Economics* 6: 117-132.

Wilt GJ van der (1994) Cost-effectiveness analysis of health care services, and concepts of distributive justice. *Health Care Analysis* 2: 296-305.
Wolfson AD, Sinclair AJ, Bombardier C, McGeer A (1982) Preference measurements for functional status in stroke patients: interrater and intertechnique comparisons. In: Kane RL, Kane RA eds. *Values and long-term care*. University of California, Lexington: 191-214.
World Bank (1993) *World development report 1993: investing in health*, New York: Oxford University Press.
Zealley AK, Aitken RCB (1969) Measurement of mood. *Proceedings of the Royal Society of Medicine* 62: 993-996.

Summary

Chapter 1 starts with a general overview of the aim of the valuation of health outcomes, with as principal elements, the quantity of life (survival) and the quality of health (health status). Until recently, the prevention of premature death formed a predominant goal of care. Now, due to the decrease of acute, life-threatening diseases, the improvement or preservation of health status has become more important. It is this shift of focus and the interest of the economic perspective that is largely responsible for the present interest in valuing health outcomes.

This thesis is directed at the issue of how to quantify health outcomes. Quality-Adjusted Life Years (QALY) models are proposed to deal with the integration of these two major elements of health to arrive at a single summary value (index) that expresses the value of (future) health of individuals or groups. This is the QALY approach. The main reason to integrate these two major elements of health into an index, is to enable comparisons such as: the health effects of different diseases, the effects of different medical interventions, the evaluation of drugs, the public's health accross different countries, et cetera. An index, for example, the Dow-Jones index, is characterized by the fact that it refers to a definite reference point. Furthermore, computational operations are allowed on its index-values. The goal of the QALY approach is to define health outcomes unambiguously and accurately.

The operationalization and valuation of health outcomes, in particular the health status element, is complicated and consist of five steps (see: Figure 1). Health status is generally thought to comprise at least three domains: physical, psychological and social functioning. Based on such a concept, a broad range of different *hypothetical* health-state scenarios can be constructed (stage II). Subsequently, such scenarios can be valued by some relevant panel using special valuation methods (stage III). We can then obtain values that represent the relative desirability of a particular health state. Subsequently, these values can be combined

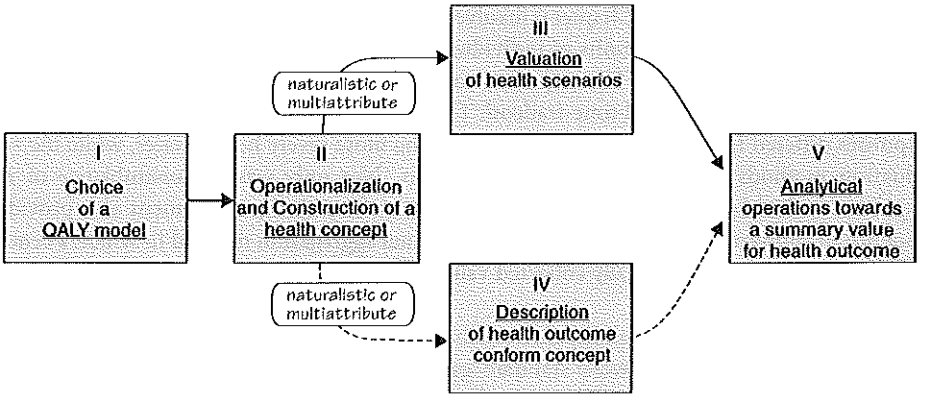


Figure 1
The five-stage approach to obtain summary values for health outcome

with survival data by performing a QALY analysis (stage V).

The QALY approach is used to identify optimal choices within the health care system, both from the viewpoint of patients and care-providers under the assumption of global consensus about the valuation of health outcomes. "Choices" refer to treatment choices at the individual (patient) level, or to health (services) policy choices at the societal level. Costs may be combined with the QALY approach in order to derive cost-effectiveness ratios: cost per QALY.

A basic requirement for the appropriate use of the QALY approach is *valid* and *reliable* quantifications for survival and health status combined. This thesis is explicitly directed to these two basic psychometric aspects of health-outcome measurement. Validity here represents the degree to which a valuation method really measures what it intends to measure, while reliability deals with the precision or reproducibility of measurements. Some key assumptions of the conventional QALY model are adressed in this thesis.

Chapters 3 – 9 are all based on empirical studies. Chapters 3 – 5 are based on a research project entitled "Standardization in Medical Technology Assessment", Chapter 6 is based on the research project "Disability Weights for Diseases in The Netherlands". Chapters 7 – 9 are based on the research project "Treatment Choice in Laryngeal Cancer". The first two studies are focused on the use of health-outcome valuation for societal (health care) applications, the third study is focused on application consequences at the individual (clinical) level.

Chapter 2 provides a conceptual overview of the QALY approaches that are used to arrive at integrated or summary values for health. First, the different applications and purposes of the QALY approach in general are discussed. Within the QALY approach, five subsequent different stages are distinguished (see: Figure 1), each representing a distinct step towards a summary value for a health outcome. Subsequently, the stages are discussed in more detail.

The next four chapters are based on valuation experiments with healthy subjects. These methodological studies were primarily directed at the various determinants of societal-based values for health outcomes. For the experiments of Chapter 3 – 5, thirteen health-state scenarios constructed according to the EuroQol (EQ-5D) health-status classification system were valued by 104 students. Chapter 6 involved judgments of 87 subjects on 18 health states.

Chapter 3 compares the results of individual responses with group responses collected by an interactive voting system. The responses consisted of valuations elicited by the valuation methods standard-gamble and time-tradeoff. The average and median values of the individual responses differed from the group values only for the more severe health states, while the two valuation methods showed almost identical results. Not only was the resemblance of the responses evoked by the two valuation methods reconfirmed, the experiment also provided evidence in support of the validity of the conventional aggregation method in which a central tendency measure of individually collected values is regarded as a representation of the study population's preference.

A detailed study on the validity and reliability of five valuation methods is presented in *Chapter 4*. The five methods were: standard-gamble (SG), time-tradeoff (TTO), rating scale (RS), willingness-to-pay (WTP) and the paired comparisons (PC) method. The multitrait-multimethod (MTMM) methodology was among others used to investigate construct validity. By transformation of the mean values using power functions the

comparability across the methods was also explored. Due to the the specific properties of the stimuli (multiattribute descriptions) the PC method proved to be not applicable as a valuation method, whereas due to the type of measurement (stimulus-scaling task), the MTMM methodology proved not to be suitable as an analytical tool. The reliability of the four other methods was satisfactory (RS = 0.77, TTO = 0.65, SG = 0.58), except for the WTP method (0.49). After a simple power transformation, the values elicited by the RS were close to the SG and the TTO values.

The experiment described in *Chapter 5* studied the so-called additivity assumption of the common QALY model. This assumption is vital to the current use of QALYs, and requires different sequences of health status in a health profile to yield similar summary values (e.g., QALYs). By using two operationalizations of the time-tradeoff method, the health-state scenarios were valued. Even after accounting for discounting effects, the results showed that at the group level a small but distinct overall effect of the sequences of the tradeoffs was present. At the individual level, we observed a small group of respondents that preferred the best health period first, though the majority of the respondents were indifferent to the sequence. The results suggest that a health profile comprising multiple levels of health status (episodic health profile) may not invariably be regarded as a chain of independent separately valued and separately discounted health profiles. This empirical finding confirms that sequences may play a role in the assessment of non-constant health profiles.

The objective of the experimental study of *Chapter 6* was to examine the effect of the extension, and inherently the refinement, of a commonly used standard multiattribute health-status classification system (EQ-5D) with an additional cognitive attribute (EQ-5D+C). It was hypothesized that by adding the cognitive attribute, the content validity of the health concept would improve, maybe at the price of decreasing reliability. Furthermore, we tested whether a simple additive model without interactions would be sufficient to predict the EQ-5D and the EQ-5D+C system. The addition of the cognitive attribute generated systematically different values compared to the standard EQ-5D version. As expected, the validity of the EQ-5D+C version was increased due to the addition of the cognitive attribute. Both systems yielded reasonably reliable valuations for the set of scenarios (EQ-5D = 0.78, EQ-5D+C = 0.74). In order to predict our non-aggregated data (i.e., individual data) a multiplicative model with interaction terms seems more appropriate.

The three following chapters are based on the assessments provided by 24 former laryngeal cancer patients and 24 lay volunteers, who valued the two treatment modalities for T3 laryngeal cancer, radiation therapy and radical surgery. The valuation tasks were repeated several times, on different occasions according to a rigorous design. Three standardized valuation methods, SG, TTO and a visual analogue scale (VAS), were used. Moreover, two different QALY models were applied: the analytical approach as applied in clinical decision analysis (Partially Decomposed QALY Model) and a holistic approach (Single Profile QALY Model). Scenarios for the decomposed model solely contained information about the stable end-states after treatment, whereas the comprehensive holistic scenarios comprised relevant information about the complete treatment course and the varying health conditions.

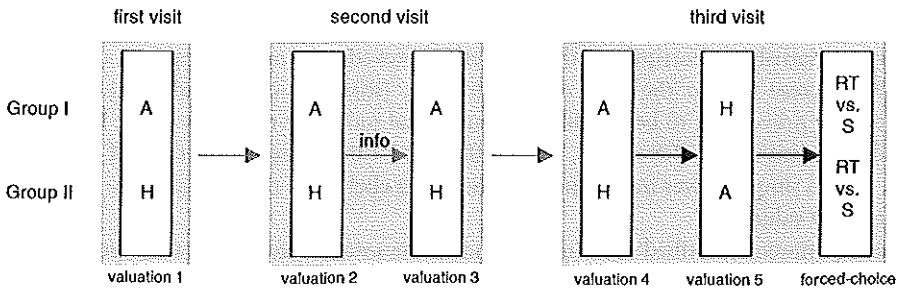


Figure 2
Schematic representation of the design of the project "Treatment Choice in Laryngeal Cancer"; A = analytical approach, H = holistic approach, RT = radiation therapy, S = surgery (Chapters 7 – 9)

Chapter 7 describes the effect of the availability of clinical information on preferences. Valuations for radiation therapy and surgery were elicited twice. In between, detailed clinical information was additionally offered about the treatment regimes (results in the thesis are presented for the TTO method only). At the group level, the offering of additional clinical information affected the values obtained by the decision analysis approach much more than the holistic approach. At the individual level we observed that a substantial proportion of the individuals changed their preferences due to the additional clinical information. This was particularly true for the decision analysis approach. The only systematic difference between the two groups of respondents was that the former patients valued both RT and S substantially higher (better) than the lay volunteers.

In *Chapter 8* the standard analytical clinical decision approach (Partially Decomposed QALY Model) was compared, under controlled conditions, with an experimental holistic approach (Single Profile QALY Model). For the analytical approach, the individual values for the optional treatment outcomes were *merged* with probability and survival data to compute Quality-Adjusted Life Expectancies, summary values analogous to QALYs. For the holistic approach, preferences for both treatments were obtained *directly* by the assessment of comprehensive scenarios. Additionally, forced-choices were obtained after completing the assessments for the two different scenarios. Average group response was similar for *both* QALY models. However, at the individual level, the number of preferences for one of the two treatments varied substantially between the two QALY models. The correspondence between treatment preferences of individuals in the analytical approach and their final forced-choice after being fully informed was high for both treatments, which suggests that the validity of the analytical approach was reasonably sustained.

The stability of valuations was examined in *Chapter 9* by applying four replications (valuation 1 – 4) of the valuation tasks (only the assessments of the analytical approach were studied). Detailed relevant clinical information was offered between valuation 2 and 3. Assessments were obtained for all three methods. Rather large instabilities could be observed across the four valuations. Values obtained at the first visit proved to be unstable, at the group level and at the individual level. Between the first and the second valuation, considerable

differences were observed for surgery, in particular for the former patients. The offering of information (after interview 2) affected the valuations of both groups. However, at the last valuation the values of the former patients and the lay volunteers became comparable for all three methods. This suggests that, provided that people are well informed, the composition of a panel becomes irrelevant. In this study, SG proved to be the most stable method for individual assessments. Replication of the rather complex valuation tasks and the rehearsal of the offered information seems necessary to elicit reliable utilities.

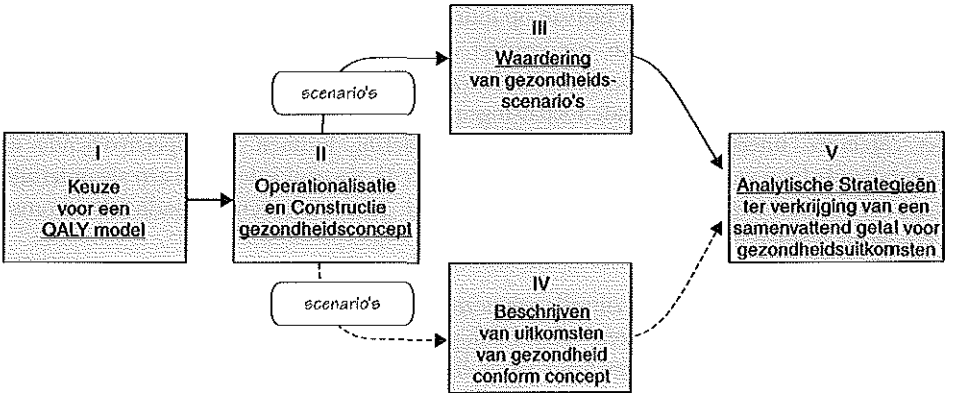
Finally, *Chapter 10* provides a general discussion about the main topics dealt with in this thesis. The QALY approach in general is discussed, especially the element of the model that has received less attention until now, i.e., the valuation of life-expectancy/survival and the combination rule to merge the valuation of health status (Q) with life expectancy/survival (L). All the relevant aspects related to the valuation of health status are thoroughly discussed. The conceptual differences between the individual and societal QALY approach and the consequences arising from these differences are discussed in a separate section. We conclude this thesis with a discussion about the future perspective of the quantification of health.

Samenvatting

Hoofdstuk 1 geeft een algemeen overzicht over het doel van het waarderen van gezondheidsuitkomsten, met als belangrijkste elementen van deze uitkomsten de duur van het leven (levensjaren) en de kwaliteit van de gezondheid (gezondheidstoestand). Waar tot voor kort het voorkòmen van voortijdig overlijden binnen de gezondheidszorg voorop stond, is meer recent, onder andere als gevolg van een drastische vermindering van het aantal manifestaties van acute, levensbedreigende ziekten, daaraan expliciet toegevoegd de verbetering of het behouden van de gezondheidstoestand. Het is deze veranderde kijk op gezondheid tezamen met de toename van aandacht voor de doelmatige inzet van financieële middelen (economisch perspectief), die hebben gezorgd voor de hedendaagse aandacht voor het waarderen van gezondheidsuitkomsten.

Dit proefschrift is gericht op de vraag of men zulke gezondheidsuitkomsten goed kan kwantificeren. Maten zoals Kwaliteits-gecorrigeerde Levensjaren (Engelse afkorting: QALY's) worden voorgesteld om de twee belangrijkste elementen van gezondheid samen te vatten tot een enkel getal (index) dat de waardering uitdrukt voor (toekomstige) gezondheid van individuen of groepen. Dit wordt wel de QALY benadering genoemd. De hoofdreden om deze twee belangrijkste elementen van gezondheid samen te willen brengen en gezamenlijk uit te drukken in een index is om te kunnen vergelijken: verschillende ziekten, effecten van verschillende medische ingrepen, effecten van geneesmiddelen, volksgezondheid van verschillende landen et cetera. Een index, vergelijkbaar met de AEX-index voor effecten, wordt gekenmerkt door het feit dat deze refereert naar een eenduidig en exact referentiepunt en daarbij dat berekeningen op de index-cijfers zijn toegestaan.

De operationalisatie en waardering van gezondheidsuitkomsten, in het bijzonder die van het element gezondheidstoestand (zie boven), zijn gecompliceerd. Het hele proces omvat vijf



Figuur 1
De vijf-fasen benadering ter verkrijging van index-waarden voor gezondheidsuitkomsten

stappen (zie: Figuur 1). In het algemeen wordt er van uitgegaan dat een gezondheidstoestand bestaat uit minimaal drie inhoudelijke domeinen, namelijk: fysiek, psychologisch en sociaal functioneren. Gebaseerd op zo'n concept kan een breed scala verschillende hypothetische gezondheidstoestandbeschrijvingen worden geconstrueerd (stadium II). Vervolgens kunnen deze beschrijvingen worden gewaardeerd door een nader te bepalen panel met gebruik van speciale waarderingsmethoden (stadium III). Hiermee verkrijgen we waarden die de relatieve preferentie van die gezondheidstoestanden uitdrukken. Vervolgens worden deze waarden samen met overlevingsgegevens ingepast in een QALY analyse (stadium V).

De QALY benadering kan worden gebruikt om optimale keuzen binnen de gezondheidszorg te identificeren, vanuit zowel het gezichtspunt van patiënten als dat van aanbieders van gezondheidszorg, onder de aanname van globale overeenstemming tussen beide groepen wat betreft de waardering van de gezondheidsuitkomsten. Keuzen kunnen verwijzen naar behandelingskeuzen op individueel (patiënt) niveau, maar evenzeer naar keuzen bij gezondheidsbeleid op maatschappelijk niveau. Overigens kunnen kosten gecombineerd worden met de QALY benadering resulterend in kost-effectiviteitsratio's: de kosten per QALY.

Voor een juist gebruik van de QALY benadering dienen waarderingen voor levensduur-gezondheidstoestandcombinaties valide en betrouwbaar te zijn. Met het psychometrische begrip validiteit bedoelen we de mate waarin een waarderingsmethode echt meet wat het behoort te meten, terwijl het begrip betrouwbaarheid betrekking heeft op de precisie en reproduceerbaarheid van de metingen. Enkele sleutelaannames van het standaard QALY model komen aan de orde in dit proefschrift.

De hoofdstukken 3 tot en met 9 zijn gebaseerd op empirische studies. Hoofdstuk 3 tot en met 5 zijn gebaseerd op een onderzoeksproject getiteld "Standaardisatie in Medical Technology Assessment", Hoofdstuk 6 is gebaseerd op het onderzoeksproject "Wegingsfactoren voor Ziekten in Nederland", terwijl de Hoofdstukken 7 tot en met 9 gebaseerd zijn op het onderzoeksproject "Behandelingskeuzen voor Larynx Kanker". De eerste twee studies zijn gericht op het gebruik van waarderingen van gezondheiduitkomsten voor maatschappelijke toepassingen (gezondheidszorg), de derde studie is gericht op de consequenties van het toepassen hiervan op individueel (patiënt of klinisch) niveau.

Hoofdstuk 2 presenteert een conceptueel overzicht van de in gebruik zijnde QALY benaderingen om tot geïntegreerde en samenvattende getallen voor gezondheid te komen. Allereerst worden de verschillende toepassingen en doeleinden van de QALY benadering in het algemeen uiteengezet. Voor de QALY benadering kunnen vijf opeenvolgende en verschillende fases worden onderscheiden (zie: Figuur 1), waarbij iedere fase een afzonderlijke stap representeert op weg naar een samenvattend getal voor een gezondheidsuitkomst. Vervolgens worden deze fases in detail besproken.

De volgende vier hoofdstukken zijn alle gebaseerd op experimenteel waarderingsonderzoek bij gezonde personen. Het waren methodologische onderzoeken, primair gericht op determinanten van waarderingen voor gezondheidsuitkomsten vanuit een maatschappelijk perspectief. Voor de experimenten van de hoofstukken 3 tot en met 5 werden dertien gezondheidstoestand-scenario's geconstrueerd door gebruik te maken van het EuroQol

(EQ-5D) gezondheidstoestand classificatie systeem. Deze werden door 104 studenten gewaardeerd. Hoofdstuk 6 heeft betrekking op een ander onderzoek waarin 18 verschillende gezondheidstoestanden door 87 personen beoordeeld werden.

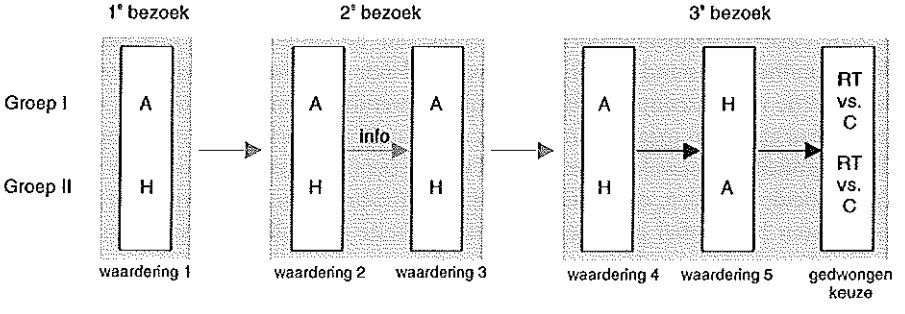
Hoofdstuk 3 vergelijkt de resultaten van individuele respons met de groepsrespons van dezelfde individuen. Deze laatste respons is verkregen met behulp van een interactief stemsysteem. De respons bestond uit waarderingen die verkregen waren met de waarderingsmethoden standard-gamble en time-tradeoff. De twee waarderingsmethoden lieten vrijwel gelijke uitkomsten zien, alleen voor de slechtere gezondheidstoestanden verschilden de gemiddelden en medianen van de waarderingen van de individuele responses met de groepswaarderingen licht. Niet alleen werd de overeenkomst van beide waarderingsmethoden bevestigd, het experiment leverde eveneens bewijs voor de validiteit van de conventionele aggregatiemethode waarbij van individuele waarderingen de centrale tendentie representatief wordt gesteld voor de waarderingen van de studie populatie.

Een studie van de validiteit en betrouwbaarheid van vijf waarderingsmethoden wordt gepresenteerd in *Hoofdstuk 4*. Deze vijf vergeleken methoden waren: standard-gamble (SG), time-tradeoff (TTO), rating scale (RS), willingness-to-pay (WTP) and the paired comparisons (PC) methode. Met de multitrait-multimethod (MTMM) methodologie is onder andere de construct-validiteit onderzocht. Tevens werden door middel van power-functies de gemiddelde waarden van de verschillende methoden op elkaar geprojecteerd. Als gevolg van de karakteristieke eigenschappen van de stimuli (multi-attribute beschrijvingen) bleek PC niet geschikt als waarderingsmethode, terwijl als gevolg van het type meten (stimulus-schaling taak), de MTMM methodologie niet toepasbaar bleek te zijn als analyse methode. De betrouwbaarheid van drie van de vier methoden was bevredigend (RS = 0.77, TTO = 0.65, SG = 0.58), de WTP methode (0.49) niet. Na een eenvoudige power-transformatie bleek dat de waarden verkregen met de RS goed overeen kwamen met die van de SG en de TTO methoden.

Het experiment beschreven in *Hoofdstuk 5* is gericht op de zogenaamde additiviteit assumptie van het standaard QALY model. Deze assumptie is cruciaal voor de huidige toepassing van QALY's en vereist dat indien er verschillende gezondheidstoestanden na elkaar voorkomen in een gezondheidsprofiel, de volgorde van deze gezondheidstoestanden geen invloed heeft op de uiteindelijke samenvattende waarden (uitgedrukt in QALY's). Met gebruik van twee verschillende operationalisaties van de time-tradeoff methode werden dezelfde gezondheidstoestand-scenario's in verschillende volgorde-profielen gewaardeerd. Zelfs na correctie voor disconteringseffecten (tijdpreferentie) lieten de uitkomsten zien dat er op groepsniveau een klein maar separaat effect was betreffende de waarderingen voor de twee gezondheidsprofielen. Hoewel de meerderheid van de respondenten geen voorkeur vertoonde voor één van beide profielen, constateerden wij op individueel niveau dat een klein aantal van de respondenten het profiel prefereerde waarin de beste gezondheidstoestand in de eerste periode voorkwam. De resultaten gaven aan dat een gezondheidsprofiel bestaande uit meerdere gezondheidstoestanden van ongelijke ernst niet onder alle omstandigheden kunnen worden beschouwd als een reeks van onafhankelijke, apart te waarderen en afzonderlijk disconteerbare eenheden. Deze empirisch bevinding benadrukt de rol die volgordes kunnen spelen in de beoordeling van gezondheidsprofielen waarbij de gezondheidstoestand varieert in de tijd.

Het doel van de experimentele studie in *Hoofdstuk 6* was het bestuderen van het effect van de uitbreiding, en daarmee verfijning, van een reeds veelvuldig gebruikt gezondheidstoestand classificatie systeem (EQ-5D) met een extra cognitie-attribuut (EQ-5D+C). De aanname was dat door het toevoegen van dit cognitieve attribuut, de inhoudsvaliditeit van het onderliggende gezondheidsconcept zou toenemen, mogelijk ten koste van de betrouwbaarheid. Daarnaast werd nagegaan of een simpel additief model zonder interacties toereikend zou zijn voor de predictie van de waarden voor EuroQol-gezondheidstoestanden met het EQ-5D en het EQ-5D+C systeem. Het bleek dat de toevoeging van het cognitie-attribuut systematisch andere waarden genereerde in vergelijking met de oorspronkelijke standaard EQ-5D versie. Zoals verwacht was de validiteit van de EQ-5D+C versie toegenomen door de toevoeging van het cognitie-attribuut. Beide systemen leverden voldoende betrouwbare waarderingen op (EQ-5D = 0.78, EQ-5D+C = 0.74). Om de niet-geaggregeerde data (d.w.z. individuele data) te voorspellen bleek een model met interactie-termen het meest geëigend.

De drie volgende hoofdstukken zijn gebaseerd op beoordelingen verricht door 24 voormalige stembandkanker (larynx-carcinoom) patiënten en door 24 leken. Zij voerden waarderingsopdrachten uit die in verband stonden met de twee behandelingsmodaliteiten voor T3-stembandkanker, namelijk bestralingstherapie en radicale chirurgische verwijdering. De waarderingsopdracht werd verscheidene malen herhaalt op verschillende momenten volgens een zorgvuldig uitgevoerde onderzoeksopzet (zie: Figuur 2). Drie standaard waarderingsmethoden, te weten SG, TTO and een visueel analoge schaal (VAS), werden gebruikt. Bovendien werden twee verschillende QALY modellen gebruikt: de analytische benadering zoals die wordt toegepast binnen de klinische besliskunde (Partieel Componenten QALY Model) en een holistische benadering (Enkelvoudig-Profiel QALY Model). Scenario's



Figuur 2
Schematische weergave van het onderzoekdsopzet van het onderzoeksprogramma "Behandelingskeuzen voor larynx kanker"; A = analytische benadering, H = holistische benadering, RT = radiotherapy, C = chirurgie (Hoofdstuk 7 – 9)

voor het Partieel Componenten Model bevatte alleen informatie over de stabiele eindtoestand na behandeling, terwijl de samengestelde holistische scenario's ook informatie bevatten aangaande het beloop (gezondheidstoestanden) en het gehele behandelingsproces.

Hoofdstuk 7 beschrijft het effect van de beschikbaarheid van klinische informatie op de waardering van gezondheidstoestanden. Waarderingen voor bestralingstherapie en chirurgische therapie werden twee maal gemeten. Tussendoor werd gedetailleerde klinische informatie aangeboden omtrent de behandelingsregimes (alleen de TTO resultaten worden in het proefschrift gepresenteerd). Het aanbieden van extra klinische informatie had op groepsniveau veel meer effect op de waarderingen verkregen met de besliskundige benadering, dan effect op de met de holistische benadering verkregen waarderingen. Ook op individueel niveau constateerden wij dat een substantieel deel van de deelnemers hun preferenties veranderde als gevolg van de extra klinische informatie. Dit was in het bijzonder het geval voor de besliskundige benadering. Het enige systematische verschil tussen beide respondentgroepen was dat de voormalige patiënten zowel voor de bestralingstherapie als voor de chirurgie substantieel hogere waarderingen hadden dan de leken.

In *Hoofdstuk 8* wordt de standaard analytische, klinisch-besliskundige benadering (Partieel Componenten QALY Model) direct vergeleken met een experimentele holistische benadering (Enkelvoudig-Profiel QALY Model). Voor de analytische benadering werden individuele waarderingen van de verschillende behandelingsuitkomsten *gecombineerd* met kanspercentages en overlevingsdata om Kwaliteit-gecorrigeerde Levensverwachtingen, welke analoog zijn aan QALY's, te berekenen. Voor de holistische benadering werden waarderingen *direct* verkregen voor beide behandelingen door de afzonderlijke beoordeling van uitgebreide scenario's. Daarnaast werden rechtstreekse keuzes verkregen na afronding van de beoordeling aan de hand van de twee verschillende scenario's. De gemiddelde groepsrespons was vrijwel gelijk voor beide QALY modellen. Op individueel niveau echter varieerde het aantal voorkeuren voor één van de twee behandelingen substantieel tussen beide QALY modellen. De overeenkomst tussen waarderingen van individuen verkregen bij de analytische benadering en hun uiteindelijke rechtstreekse keuze voor één van de behandelingen, na volledig geïnformeerd te zijn, was hoog, hetgeen de analytische benadering redelijk ondersteunt.

De reproduceerbaarheid van waarderingen werd onderzocht in *Hoofdstuk 9* door vier herhalingen (waardering 1 – 4) van de waarderingsopdracht (enkel de beoordelingen van de analytische beandering werden onderzocht). Gedetailleerde, relevante klinische informatie werd aangeboden tussen de tweede en derde waardering. Beoordelingen werden verkregen met drie waarderingsmethoden. Een behoorlijke mate van instabiliteit kon worden waargenomen over de vier waarderingen. In het bijzonder waarderingen verkregen tijdens het eerste bezoek bleken instabiel te zijn, zowel op individueel als op groepsniveau. Tussen de eerste en de tweede waardering werden behoorlijke verschillen zichtbaar voor de chirurgische therapie, in het bijzonder voor de voormalige patiënten. Het aanbieden van de informatie had ook een effect op de waarderingen van de twee groepen. Echter, tijdens het laatste bezoek werden de waarderingen voor de voormalige patiënten en de leken vergelijkbaar voor alle drie de waarderingsmethoden. Dit duidt erop dat mits mensen goed geïnformeerd zijn, de samenstelling van een beoordelaarsgroep irrelevant wordt. De SG methode bleek in deze studie de meest stabiele methode te zijn voor individuele beoordelingen. Herhaling van

tamelijk gecompliceerde beoordelingsopdrachten zoals verricht in deze studie, alsmede het herhaald aanbieden van informatie lijkt een voorwaarde te zijn voor het verkrijgen van betrouwbare waarderingen.

Hoofdstuk 10 tenslotte bediscussieert de hoofdpunten van dit proefschrift. Het algemene QALY model komt aan de orde, vooral het element van het model dat zo weinig aandacht heeft gekregen tot nu toe: de waardering van overlevings-/levensjaren en de combinatieregels die waarderingen voor gezondheidstoestanden (Q) met overlevings-/levensjaren (L) verbinden. In dit hoofdstuk worden diverse aspecten besproken die de waardering van gezondheidsuitkomsten beïnvloeden. De conceptuele verschillen en praktische consequenties tussen de individuele en de maatschappelijke QALY benadering zijn onderwerp in de derde sectie. Wij besluiten dit proefschrift met een discussie over het perspectief voor het kwantificeren van gezondheid.

Co-autors

G.J. Bonsel, MD, PhD
Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, Amsterdam, The Netherlands

J. van der Donk, PhD
Dr. Daniel den Hoed Cancer Center, Rotterdam, The Netherlands

M.L. Essink-Bot, MD, PhD
Department of Public Health, Erasmus University Rotterdam, Rotterdam, The Netherlands

J.D.F. Habbema, PhD
Department of Public Health, Erasmus University Rotterdam, Rotterdam, The Netherlands

P.C. Levendag, MD, PhD
Dr. Daniel den Hoed Cancer Center, Rotterdam, The Netherlands

C.A. Meeuwis, MD, PhD
Dr. Daniel den Hoed Cancer Center, Rotterdam, The Netherlands

P.I.M. Schmitz, PhD
Dr. Daniel den Hoed Cancer Center, Rotterdam, The Netherlands

M.E.A. Stouthard, PhD
Institute of Social Medicine, University of Amsterdam, Academic Medical Centre, Amsterdam, The Netherlands

List of Scientific Publications

Agt HME van, Essink-Bot ML, Krabbe PFM, Bonsel GJ (1994) Test-retest reliability of health status valuations collected with the EuroQol questionnaire. *Social Science & Medicine* 39: 1537-1544.

Essink-Bot ML, Royen L van, Krabbe PFM, Bonsel GJ, Rutten FFH (1995) The impact of migraine on health status. *Headache* 35: 200-206.

Essink-Bot ML, Krabbe PFM, Agt HME van, Bonsel GJ (1995) NHP or SIP – a comparitive study in renal insufficiniency associated anemia. *Quality of Life Research* 5: 91-100.

Essink-Bot ML, Krabbe PFM, Bonsel GJ, Aaronson NK (1997) An empirical comparison of four generic health status measures: the Nottingham Health Profile, the MOS 36-item Short-Form Health Survey, the COOP/WONCA charts, and the EuroQol instrument. *Medical Care* 35: 522-537.

Krabbe PFM, Dijksterhuis GB (1992) Toetsing bij de driehoeksmethode. *Voedingsmiddelentechnologie* 19: 19-23.

Krabbe PFM, Essink-Bot ML, Bonsel GJ (1996) On the equivalence of collectively and individually collected responses: standard-gamble and time-tradeoff judgments of health states. *Medical Decision Making* 16: 120-132.

Krabbe PFM, Essink-Bot ML, Bonsel GJ (1997) The comparability and reliability of five health-state valuation methods. *Social Science & Medicine* 45: 1641-1652.

Krabbe PFM, Bonsel GJ (1998) Sequence effects, health profiles and the QALY model: in search of realistic modeling. *Medical Decision Making* 18: 178-186.

Punter PH, Krabbe PFM, Dijksterhuis GB (1992) Multivariate data analyse in voedingsmiddelenonderzoek. *Voedingsmiddelentechnologie* 12: 18-21.

Tuiten A, Panhuysen G, Everaerd W, Koppeschaar H, Krabbe P, Zelissen P (1993) The paradoxical nature of sexuality in anorexia nervosa. *Journal of Sex & Marital Therapy* 19: 259-275.

Tuiten A, Panhuysen G, Kopeschaar H, Fekkes D, Pijl H, Fröhlig M, Krabbe P, Everaerd W. (1995) Stress, serotonergic function, and mood in users of oral contraceptives. *Psychoneuroendocrinology* 20: 323-334

Glosssary

General aspects

Attribute — Technical use: a specific domain of a health-status classification (e.g., physical, psychological, social functioning).

DALY (Disability-Adjusted Life Years) — A two-attribute summary value, analogous to QALYs, that combines the number of life years (L) lost due to premature death and the number of years lived with disability (health status: Q). DALYs are conceptually identical to QALYs lost, and like the latter need some reference or comparison state.

Dimension — See *domain.*

Discounting — Technical use: a computation procedure to adjust past, current and future monetary gains and losses *as if* all transactions occurred at some defined point in time (t_0). In decision-analytic applications t_0 is usually chosen to reflect the instant a decision has to be made. If t_0 is "now", the procedure yields so-called present value of all gains and losses. Although developed for investment support to account for intertemporal monetary effects, the method of discounting is also applied to health to express future outcomes in terms of their present value. The usual approach is $\left(\frac{1}{1+r}\right)^{n-1}$, where n is the n-th life year. and r reflects the disount rate. Parameter r is comparable to the (compound) interest rate.

Domain — A specific aspect of health status (e.g., physical, psychological, social functioning). Technical use: see *attribute.*

Duration — Used in the context of a period or QALY-ple. See *survival.*

Ex Ante — A decision situation viewed from *before* an event occurs, or an action to be undertaken, before the outcome is known. Uncertainty plays a role in judging such situations. If the situation deals with health interventions, ex ante judgment of the health effects and other effects at the individual level have to account for the (dis)utility of the uncertainty of outcomes. This can be managed by including probability-related adjustments and/or by applying aggregational rules other than expected utility.

Ex Post — A situation viewed *after* an even occurs, or an action has taken place or an outcome in known. Uncertainty is irrelevant here, as opposed to ex ante health situations.

Health — This concept has multiple interpretations, e.g., well-being in as far dependent on the presence of symptomatic diseases or knowledge of having contracted a disease. Technical use: (expected) health-related quality and (expected) quantity of life related to a specific reference point in time.

Health Description — See *scenario.*

Health Element — Technical use: conceptual part of a QALY model [health status (Q), life years (L)].

Health Indicator — Aspect related to health by which health outcome as the measure of effectiveness of health services can be comprehensively defined (e.g, health status, survival, clinical status). The health elements (see above) are a special case.

Health Outcome — Technical use: health following a decision and/or specific intervention, that is expressed by a set of health-related indicators (health status, survival).

Health-Related Quality-of-Life (HRQoL) — See *health status.*

Health State — A period of (assumed) constant health status. Technical use: health description of any degree of disaggregation serving as a (descriptive) stimulus in a valuation task (timeless concept).

Health-State Description — See *health-state scenario.*

Health-State Scenario — Technical use: specification of a health outcome by means of a description of health status or a specific classification for health status (Q), eventually also comprising the QALY element survival (L) and other health indicators.

Health-State Valuation — Technical use: the empirical (direct) assignment of a numerical value to a specified (element of) health outcome. Can also be performed by so-called indirect valuation or scaling methods (e.g., paired comparisons method).

Health Status — Technical use: quality of health expressed in non-normative, descriptive terms, where duration of survival is ignored (timeless concept).

Life Expectancy — The number of years a person is ecxpected to live, given a particular age.

Life Years — Unit of measurement to express the health element survival/duration.

Multiattribute Health-Status Classification System — A tool to formally describe health status; it requires health to be operationally defined according to several domains with hierarchically ordered levels per domain. The actual health status of the patients can be scored using the pre-established collected preference weights for all combinations of scoring on the different domains and their levels, eliminating the need to undertake primary measurement of preferences.

Multiattribute Scenario — A health-status scenario constructed on the basis of a multiattribute health-status classification sytem.

Naturalistic Health-Status Classification System — A tool to describe health status; it requires health to be operationally defined according to several domains. In contrast to the *multiattribute health-status classification system*, a formal structure is absent and no hierarchical levels (e.g., quantifications) are associated with each level.

Naturalistic Scenario — A health-status scenario based on a naturalistic health-status classification system, usually operationalized in terms of symptoms or disease states.

Preference — A numerical judgment of the desirability of a particular outcome or situation. Technical use: valuation expressed as a utility (a positive real number in the range 0–1 with ratio measurement properties). Also known as preference score, preference weight or preference value.

Probability — Technical use: number between 0 and 1 expressing the likelihood of occurrence.

Prognosis — Future outlook of a person's health, expressed in any form, subjectively or objectively. Technical use: expected developement of health status (Q) and survival (L), given some point in time.

Scenario — See: *health-state scenario.*

QALY (Quality-Adjusted Life Years) — A two-attribute summary value expressing the evaluation of a future health profile, combining preferences for two distinct health elements: health status (Q) and life years (L). Note: the measurement of the two elements is far from standaridized.

QALY Model — A weighting scheme that combines health status (Q) and life years (L) estimates arriving at a summary value for health profiles.

QALY-ple — A discrete period, uniquely defined by (i) constant health status (Q),(ii) life years (L) expressed in numerical, semi quantitative or qualitative terms, and eventually (iii) a specific end-point (c).

QALY-ple Chain — A succession of two or more QALY-ples.

QALY-ple Tree — Two or more lifetime health profiles (e.g., QALY-ple chains). Such a tree represents the possible or realized life prospects of one or more individuals.

Quality of Life — A broad construct reflecting subjective or objective judgments concerning all (generic) or selected (domain-related, disease-related) aspects of an individual's existence. These aspects may include health, economic, political, cultural, environmental, aesthetic, and spiritual aspects. If the focus is restricted to health only it is referred to as Health-Related Quality of Life (HRQoL).

Response Mode — Technical use: the specific operations and conditions under which responses (valuations) are collected.

Risk Attitude — Technical use: a specific, individual characteristic that affects the individual valuation of a life prospect under uncertainty. Special case: Pliskin's QALY model specifies risk attitude as parameter r.

Stimulus Mode — The specific operations and conditions used to offer the stimulus (health-state scenarios) to be valued.

Summary Value — A composite measure combining unrelated attributes. In the context of health it is often referred to as a QALY. Synonyms: summary measure, aggregate health measures, (summary) index, morbidity-mortality value.

Survival — One of the health indicators, and technically one of the health elements. Technical use: in epidemiology the (expected) number of life years for a group or an individual, usually after exposure or a medical intervention. End-points of survival periods can be indicated by three possible states: dead, alive or so-called censored (unspecified) states.

Unit of Measurement — The smallest possible numerical expression of an element of health outcome. In most applications determined by the opeationalization of the valuation method used (for example: standard-gamble; alteration of the probability steps).

Health representations in valuation studies

Episodic Health Profile — A (future) health profile, comprising multiple health states, and *not* necessarily ending with death.

Health Profile — A representation of a specified time period comprising one or more health states (combination of QALY-ples).

Life Prospect — Technical use: expectation of lifetime health profiles.

Lifetime Health Profile — (Future) health profile, usually comprising multiple health states, ending with death.

QALY-ple — A discrete period, uniquely defined by (i) constant health status (Q),(ii) life years (L) expressed in numerical, semi quantitative or qualitative terms, and eventually (iii) a specific end-point (c).

Measurement aspects

Analytical Approach — A formal approach to derive the valuation of health outcomes by computational strategies based on the imputation of (individually) elicited values. Generally combined with empirically or theoretically obtained estimates from data bases or literature.

Assessment — Any formal measurement procedure resulting in a value judgment. Also refers to the measurement (data) itself.

Appraisal — The numerical quantification of a particular phenomenon. Technical use in this thesis: the implicit assessment of a particular phenomenon, in this context health outcomes. See also: *valuation.*

Decomposed Approach — Indentical to the *analytical approach.*

Elicitation Method — See *valuation method.*

Expected Utility Theory — A formal axiomatic theory for the valuation of – in this context – health outcomes under uncertainty.

Holistic Approach — An approach to derive the valuation of a health profile consisting of a combination of different health states by assessing the whole profile.

Measurement — Technical use: the process of arriving at numerical figures or categorical scores to reflect the existence (to some degree) of a particular phenomenon. In this context the technical, methodological and theoretical aspects of the description and valuation of health outcomes (e.g., information process, the validity and reliability of such values etc.).

Research Design — The specific operationalization of a study. A design is appropriate if it is capable of examing the reseach question.

Utility — Numerical value assigned to a health state, satisfying the axioms of Expected Utility Theory.

Valuation — The explicit numerical quantification of a particular phenomenon, in this context health outcomes.

Valuation Method — A method to elicit values. If the conditions of Expected Utility Theory are fulfilled then these values may be called utilities.

Valuation Task — The setting used to obtain values for, in this context, health outcomes. This requires the following: the choice of a specific valuation method; its specific operationalization; the type of scenario used to describe the health-information (naturalistic or multi-attribute); the method of offerring the information (stimulus mode); the method of collecting the valuations (response mode), etc.

Value — Technical use: numerical value assigned to a health state, not necessarily satisfying the axioms of Expected Utility Theory. General use: numerical value assigned to a health state.

The elements of the QALY models and the formal notation used in this thesis

A, a — Attributes/attribute (dimensions/dimension) of health status.

L, l — life years/life years; usually but not necessarily expressed in quantitative terms.

N, n — Number of individuals.

P, p — Probabilities.

Q, q — Health states/health state (status); may consist of attributes.

$U(...)$ — Utility (or valuation) of a specific element.

$U(Q, L)_I$, *Model I* — Fully Decomposed QALY Model.

$U(Q, L)_{II}$, *Model II* — Partially Decomposed QALY Model.

$U(Q, L)_{III}$, *Model III* — Single Profile QALY Model.

$U(\mathbf{Q}, L)_{IV}$, *Model IV* — Multiple Profiles QALY Model.

Dankwoord

Velerlei ervaringen liggen ten grondslag aan de uiteindelijke "proeve van bekwaamheid" die u hier in handen heeft. Hoezeer ook de psycholoog in mij beseft welk een stempel mijn vroege jeugd, mijn ouders en mijn schoolperiode op mij hebben gedrukt, zal ik alleen mijn saillante academische ervaringen de revue laten passeren.

Dr. J.J.A. Tuiten en dr. G.E.M. Panhuysen van de vakgroep Theorie en Grondslagen van de Psychologie (Universiteit Utrecht) dank ik voor de intensieve en prettige samenwerking die wij hadden tijdens de afronding van mijn wetenschappelijke opleiding. Mede door jullie, Adriaan en Geert, kwam ik in aanraking met wetenschappelijk *onderzoek* en werd mijn belangstelling hiervoor gewekt.

Na mijn afstuderen kreeg ik dankzij prof. dr. W.H.J. Meeuws mijn eerste wetenschappelijk betrekking. Wim, de vakgroep "Jeugd, Gezin & Levensloop" (Universiteit Utrecht) alsmede het Sociaal en Cultureel Planbureau, alwaar ik gedetacheerd was, hebben bij mij goede herinneringen achtergelaten.

Overduidelijk ben ik wetenschappelijke gevormd door twee jaar te werken bij het sensorische onderzoekbureau Oliemans Punter & Partners te Utrecht. Bij dit jonge bedrijf gebeurde veel, was veel mogelijk en werden nieuwe ideeën en ontwikkelingen niet geschuwd. Pieter, van jou heb ik geleerd de zaken goed en grondig aan te pakken en niet altijd de begane paden te volgen. Ik hoop dat hiervan iets terug te vinden is in mijn proefschrift. Ook naar dr. G.B. Dijksterhuis wil ik mijn erkentelijkheid uitspreken. Garmt, je was een oprechte en enthousiaste collega die mij heeft laten zien dat een mens zijn eigen koers kan én moet varen.

I wish to thank members of the EuroQol Group for the numerous discussions we have had on the topic of quantification of health status. Not only has the scientific debate within the Group been fruitful but membership of the Group at a social and personal level has been most enjoyable. De EuroQol Group Business Manager dr. F. Th. de Charro dank ik voor zijn hartelijkheid tijdens onze contacten. Wellicht heb je gelijk Frank, als je refereert aan onze katholieke achtergrond om onze goede verstandhouding te verklaren.

Dankbaarheid spreek ik uit voor de gelegenheid die het instituut Maatschappelijke Gezondheidszorg (Erasmus Universiteit Rotterdam) mij bood om tot afronding van dit proefschrift te komen. Na de formele voltooiing van de onderzoeksprojecten heeft dit instituut mij het laatste jaar financieel en facilitair gesteund, waarvoor ik in het bijzonder prof. dr. P.J. van der Maas en prof. dr. J.D.F. Habbema wil bedanken.

René Eijkemans, Ewout Steyerberg en Pieta Krijnen van het Centrum Klinische Besliskunde waren de laatste twee jaren op het instituut mijn naaste collega's. Jullie bedank ik voor de aangename en informele contacten die ik met jullie mocht hebben. De schone avonden die ik samen met Réne en Ewout in Turijn had zal ik niet snel vergeten. De wijn daar was uitstekend.

De dames van het secretariaat: Tineke Kurtz, Ilse Philips, Mona Richter, Aty Slikkerveer en alle anderen dank ik voor hun hulp op vele fronten, maar bovenal voor jullie hartelijkheid en de gelegenheid die ik bij jullie vond om af en toe over iets anders dan het werk te praten.

Mijn promotor prof. dr. J.D.F. Habbema wil ik ook bedanken. Dik, ondanks je vele werkzaamheden was je altijd aanspreekbaar en bewaakte je nauwgezet de kwaliteit en coherentie van dit proefschrift. Daarbij zorgen je humor en relativeringsvermogen ervoor dat samenwerken met jou een aangenaam gebeuren is.

De co-auteurs van de Dr. Daniel den Hoed kliniek, dr. J. van der Donk, prof. dr. P.C. Levendag, dr. C.A. Meeuwis en dr. ir. P.I.M. Schmitz bedank ik voor hun bijdragen aan het larynxproject en ook voor de genoegelijke momenten die wij hebben gekend.

Mijn huidige collega's van de afdeling Medical Technology Assessment (Katholieke Universiteit Nijmegen) bedank ik voor de belangstelling die zij toonden voor de voortgang van mijn proefschrift en de medewerking die ik van hen heb gekregen bij de afronding ervan.

Een speciale vermelding verdient dr. M.L. Essink-Bot. Bijna vijf jaar lang was zij mijn collega en kamergenote. Beste Marie-Louise, in onze samenwerking hebben wij denk ik veel aan elkaar gehad. Niet alleen in wetenschappelijk zin, maar ook in persoonlijke zin. Vreugde en zorgen omtrent relaties, werken, politiek (mijn besognes bij de sociaal-democraten) en de kinderen hebben wij gedeeld. Onze gezinsuitbreidingen liepen niet parallel aan de bevallingen van onze proefschriften. Maar je hielp me. In december 1995 mocht ik vanwege het tijdelijk maar dramatisch wegvallen van jouw stem, als jouw paranimf de introductie van jouw proefschrift presenteren. Een betere voorbereiding op mijn eigen promotie is schier ondenkbaar.

Mijn grootste dank gaat zonder meer uit naar mijn co-promotor dr. G.J. Bonsel. Beste Gouke, ik ken weinigen die gedreven door theoretische en praktische interesses zo betrokken zijn bij hun wetenschappelijk métier als jij. Voor mij was je altijd bereid om advies te geven en om manuscripten kritisch door te lezen. Naarmate het proefschrift vorderde hadden wij in toenemende mate inhoudelijke discussies en allengs kwam ons helderder voor ogen te staan hoe de problematiek rondom het kwantificeren van gezondheid in elkaar steekt. Dit heeft vanzelfsprekend tot extra werk geleid (zie: Appendix A). Door jouw aansturing, zeg maar gerust aanvuring, ben ik niet alleen gekomen tot een bovenmatige inspanning, maar kan ik eveneens terugblikken op een enerverende en plezierige periode.

Vera, in december 1996 hebben wij uitbundig gevierd dat wij twaalf-en-half jaar getrouwd waren. In de daaropvolgende periode moest de promotie tot een goed eind worden gebracht. Tevens moest een nieuwe baan gevonden worden, werd ik onverwachts tussentijds beëidigd tot raadslid en kocht jij ons huis. Al die uren dat ik werkte aan dit proefschrift heb je mij volop gesteund. Het leven ging gewoon door! Edoch, de voltooiing ervan was waarlijk geen sinecure en de laatste maanden werd ik terecht door jou aangespoord om tot afronding te komen. Voor wat de toekomst betreft, verlaat ik mij op de opening van een lied dat ik jou de afgelopen jaren in huis zo vaak heb horen zingen: "Sag', welch wunderbare Träume halten meinen Sinn umfangen ...".

Zeist, juni 1998

Curriculum Vitae

Paulus Franciscus Maria Krabbe is geboren op 3 februari 1959 te Heemskerk. Zijn lagere-schoolperiode bracht hij door in Haarlem. In 1976 behaalde hij zijn einddiploma HAVO te Lelystad. In datzelfde jaar begon hij in Amsterdam de opleiding grafische-MTS. Na afronding van deze opleiding in 1980 was hij enkele jaren werkzaam in de grafische branche. In de periode 1984-1986 volgde hij te Utrecht de avondopleiding VWO. Na het behalen hiervan ving hij aan met de studie Psychologie aan de Rijksuniversiteit van Utrecht. Gedurende de studie werden enkele student-assistentschappen vervuld. Drie-en-half jaar studie werd in 1990 afgerond met als hoofdrichtingen Theoretische en Klinische Psychologie alsmede Methoden & Technieken van onderzoek.

Na zijn doctoraalexamen is hij in de vorm van een samenwerkingsverband werkzaam geweest bij de Rijksuniversiteit Utrecht (vakgroep Jeugd, Gezin & Levensloop) en het Sociaal en Cultureel Planbureau. Vervolgens heeft hij twee jaar gewerkt bij het sensorisch onderzoekbureau Oliemans, Punter & Partners te Utrecht. Vanaf medio 1992 tot mei 1997 werkte hij als onderzoeker bij het instituut Maatschappelijke Gezondheidszorg aan de Erasmus Universiteit Rotterdam. Momenteel is hij werkzaam bij de afdeling Medical Technology Assessment (vakgroep Medische Informatiekunde, Epidemiologie en Statistiek) van de Faculteit der Medische Wetenschappen aan de Katholieke Universiteit Nijmegen.

Gedurende de jaren waarin dit proefschrift tot stand kwam, is hij tevens politiek actief geweest, ondermeer in de functie van raadslid. In het kader van deze politieke activiteiten zijn enkele politieke geschriften van zijn hand verschenen in onder andere landelijke dagbladen. Daarnaast houdt hij zich bezig met het grafisch verzorgen van hoogwaardig drukwerk (dit proefschrift).

De auteur is getrouwd met Divera Maria Breg. Samen hebben zij drie kinderen: Tom (1987), Oskar (1989) en Olga (1993).

Colofon

Cover design, Illustrations and Typography
Theta Research, Zeist

Desktop Publishing
Theta Research, Zeist

Fonttypes
Adobe Garamond, Helvetica, Tekton;
Adobe Inc.

Paper
Promail plus respect;
Proost en Brandt, Diemen

Printer
Elinkwijk, Utrecht

Some of the most popular typefaces in history are those based on the types of the sixteenth-century French printer, publisher, and type designer Claude Garamond (1480-1561), whose sixteenth-century types were modeled on those of Venetian printers from the end of the previous century. The Garamond typeface and its variations have been a standard among book designers and printers for centuries.
Designer Robert Slimbach went for the Adobe company to the Plantin-Moretus museum in Antwerp, Belgium, to study the original Garamond typefaces. The italics are based on types by Robert Granjon, a contemporary of Garamond.

Helvetica is one of the most popular typefaces of all time. It was designed by Max Miedinger in 1957. The design is based on the grotesques of the late nineteenth century.

Based on the hand lettering of West Coast architect Frank Ching, who wrote out the text for his books, Tekton was designed by David Siegel in 1989.